



Introduction à la FOUILLE DE TEXTES et positionnement de l'offre logicielle

Patrice Bellot (Aix-Marseille Univ, CNRS)

patrice.bellot@cnrs-dir.fr

ANF TDM — mars 2021

QU'EST-CE QUE LA FOUILLE DE TEXTES ?

Le croisement de plusieurs domaines

- L'analyse et la fouille de données (Data Mining)
- Le Traitement Automatique des Langues
- La recherche et l'extraction d'information

QUELQUES DIFFICULTÉS...

Quelle que soit la nature des données :

- Structures peu normalisées, formats variés
- Les fameux V du Big Data : Volume, véracité, variabilité, valeur, vitesse

Documents, textes et langues :

- Données hétérogènes ou multimodales
- Multilinguisme (lexiques, terminologies, syntaxes)
- La langue est ambiguë

Le Droit et les bonnes pratiques pour la mise en œuvre du TDM :

Des difficultés
génériques.

Des standards
nécessaires.

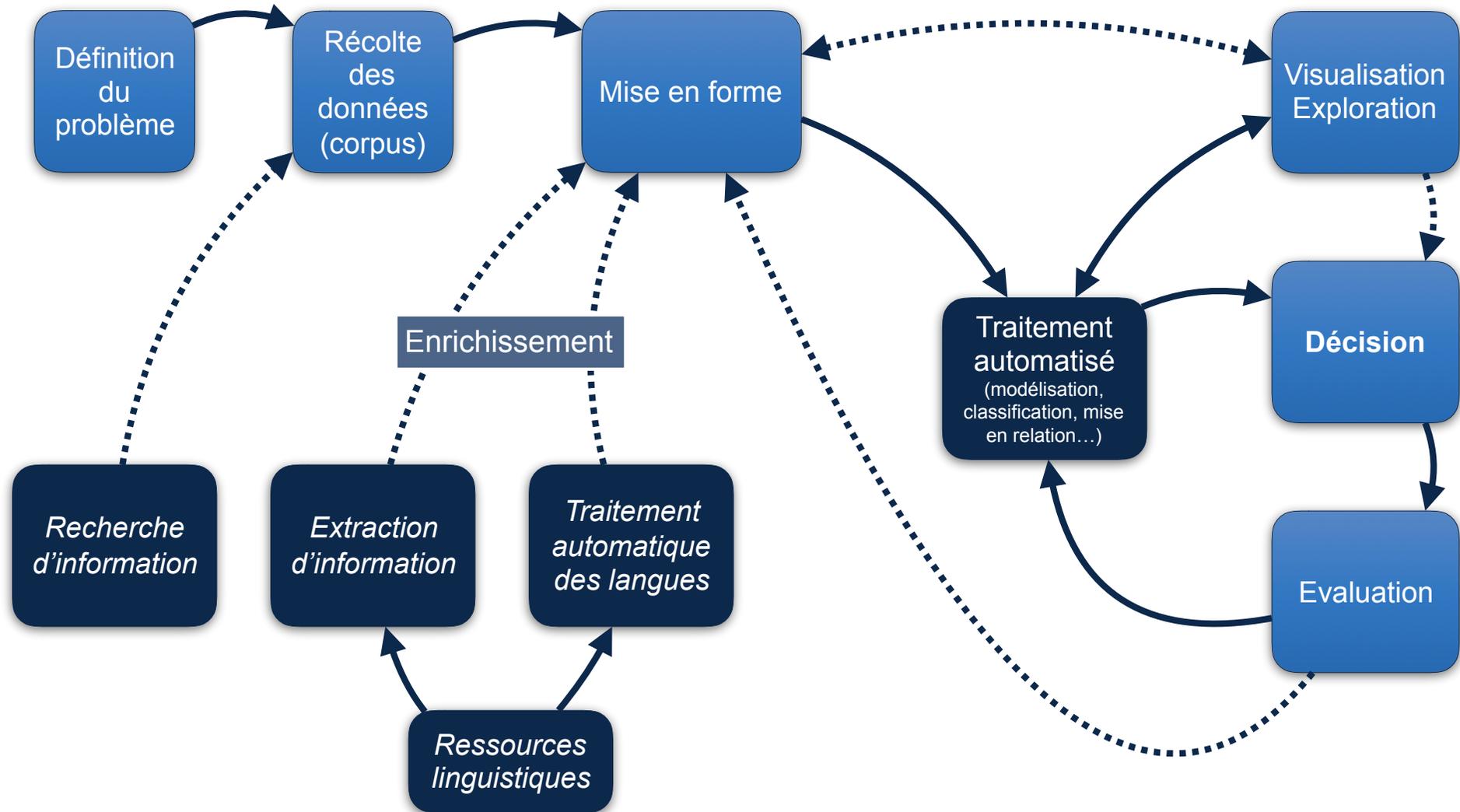
Des solutions
à partager.



*loi française Pour une République
Numérique du 7 octobre 2016 et la
directive européenne sur le droit
d'auteur dans le marché unique
numérique du 26 mars 2019*

Des données, des méta-données et des formats

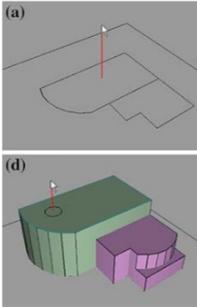
Un processus de fouille de textes



Des données

Virtual Reality (2006) 10:135-147

Fig. 3 Sketching interface. **a.** Dragging outward from a closed contour creates a volume. **c. d** Drawing a contour on a face and dragging inward sculpts the volume. **d. e** The user creates a new shape by dragging outward. **e. f** The user can also “stretch” an object by dragging the face directly. (Red lines added to visualize drag operations)



designs faster with this tool than traditional CAD systems (Oh et al. 2006b). Also, studies with naïve users (i.e. people who had never used 3D design sys-

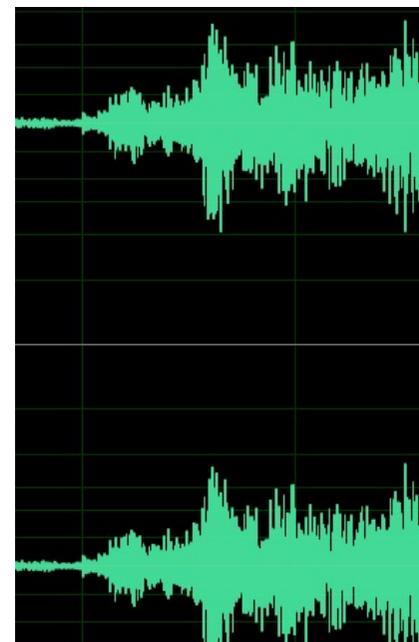
s que Gilberte invita
comme les autres a
endais s'échapper d
dans l'émotion qu
qu'elle j'allais assiste
j'atteignisse le pali
la vie antérieure et
retirer mon foulard
rder l'heure pour no
urs, tout en bois, co
ons de rapport de c
l'idéal d'Odette et
ourvu d'une pancar

chez nous, sur laquelle on lisait ces mots:
« Arriver de l'ascenseur pour descendre »
« une chose de tellement prestigieuse que
« ts que c'était un escalier ancien rapporté
« vanni. Mon amour de la vérité était si gra
« pas hésité à leur donner ce renseigne
« s su qu'il était faux, car seul il pouvait
« ir pour la dignité de l'escalier des Swan
« que moi. C'est ainsi que devant un ignor

Texte



Image



Audio

Des données « numérisables »

s, certaines des amies que Gilberte invita
nt obligées de partir comme les autres a
, dès l'escalier j'entendais s'échapper de
urmure de voix qui, dans l'émotion qu
ionie imposante à laquelle j'allais assiste
ent, bien avant que j'atteignisse le palie
attachaient encore à la vie antérieure et
souvenir d'avoir à retirer mon foulard
; au chaud et de regarder l'heure pour ne
l. Cet escalier, d'ailleurs, tout en bois, ce
dans certaines maisons de rapport de c
vait été si longtemps l'idéal d'Odette et
ôt se déprendre, et pourvu d'une pancar
chez nous, sur laquelle on lisait ces mots
rvir de l'ascenseur pour descendre »
que chose de tellement prestigieux que
ts que c'était un escalier ancien rapporté
vann. Mon amour de la vérité était si gra
pas hésité à leur donner ce renseigne
s su qu'il était faux, car seul il pouvait
ir pour la dignité de l'escalier des Swan
que moi. C'est ainsi que devant un ignor

Texte : des symboles (mots
constitués de caractères)

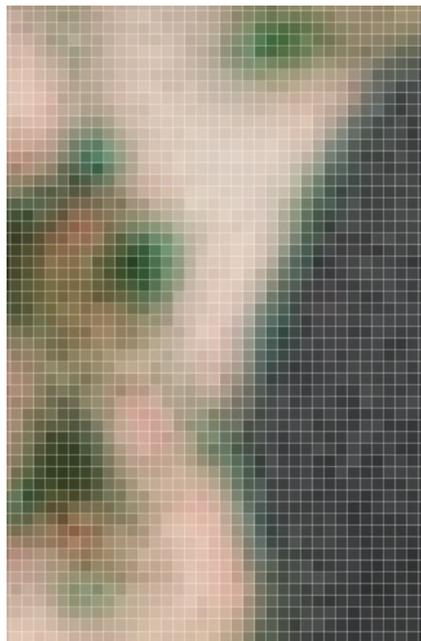
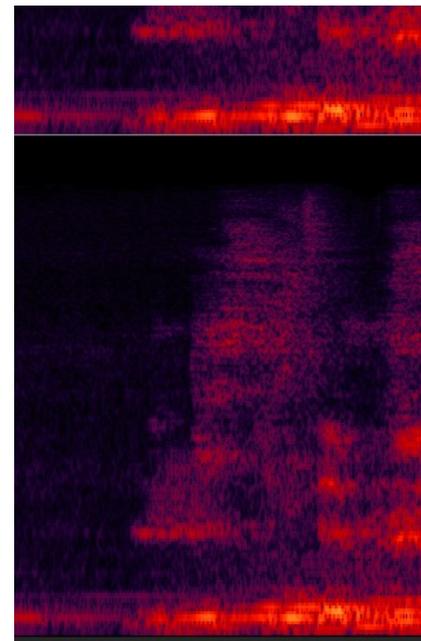


Image : des pixels



Audio : des fréquences

L'encodage des caractères

✓ Par défaut Occidental (ISO Latin 1) Occidental (Mac OS Roman)
Unicode (UTF-8)
Japonais (Shift JIS) Japonais (ISO 2022-JP) Japonais (EUC) Japonais (Shift JIS X0213)
Chinois traditionnel (Big 5) Chinois traditionnel (Big 5 HKSCS) Chinois traditionnel (Windows, DOS)
Coréen (ISO 2022-KR) Coréen (Mac OS) Coréen (Windows, DOS)
Arabe (ISO 8859-6) Arabe (Windows)
Hébreu (ISO 8859-8) Hébreu (Windows)
Grec (ISO 8859-7) Grec (Windows)
Cyrillique (ISO 8859-5) Cyrillique (Mac OS) Cyrillique (KOI8-R) Cyrillique (Windows) Ukrainien (KOI8-U)
Thaïlandais (Windows, DOS)
Chinois simplifié (GB 2312) Chinois simplifié (HZ GB 2312) Chinois (GB 18030)
Europe centrale (ISO Latin 2) Europe centrale (Mac OS) Europe centrale (Windows Latin 2)
Vietnamien (Windows)
Turc (ISO Latin 5) Turc (Windows Latin 5)
Europe centrale (ISO Latin 4) Balte (Windows)

Dec	Hex	Char	Dec	Hex	Char
64	40	@	96	60	'
65	41	A	97	61	a
66	42	B	98	62	b
67	43	C	99	63	c
68	44	D	100	64	d
69	45	E	101	65	e
70	46	F	102	66	f
71	47	G	103	67	g
72	48	H	104	68	h
73	49	I	105	69	i
74	4A	J	106	6A	j
75	4B	K	107	6B	k
76	4C	L	108	6C	l
77	4D	M	109	6D	m
78	4E	N	110	6E	n
79	4F	O	111	6F	o
80	50	P	112	70	p
81	51	Q	113	71	q

Dec	Hex	Char	Dec	Hex	Char
7	00A8	À	160	A0	À
7	00B8	ı	161	A1	ı
7	00C8	È	162	A2	È
7	00D8	ø	163	A3	ø
7	00E8	è	164	A4	è
7	00F8	ø	165	A5	ø

ASCII (années 1960, sur 7 bits)
 American Standard Code for
 Information Interchange

ISO-Latin 1 sur 8 bits
 (ISO/CEI 8859)

<https://unicode.org/emoji/charts/full-emoji-list.html>
<https://home.unicode.org>

face-smiling													
N°	Code	Browser						Appl	Goog	FB			
940	á	é	ή	ί	ύ	α	β	γ	δ	ε	☺	☺	☺
950	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο	☺	☺	☺
960	π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	☺	☺	☺
970	ϊ	ϋ	ό	ύ	ώ	ϋ	ϋ	ϋ	ϋ	ϋ	☺	☺	☺
980	ÿ	φ	ω	χ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	☺	☺	☺
990	Ϝ	ϝ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ	Ϥ	ϥ	☺	☺	☺
1000	Ϝ	ϝ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ	Ϥ	ϥ	☺	☺	☺
1010	с	ј	Ѡ	ε	э	б	р	С	М	М	☺	☺	☺
1020	р	Ѡ	ѡ	È	Ë	Ђ	Ѓ	Є	Ѕ	Ѕ	☺	☺	☺
1030	І	Ї	Ј	Љ	Њ	Ћ	Ќ	Ў	Љ	Љ	☺	☺	☺
1040	А	Б	В	Г	Д	Е	Ж	З	И	Й	☺	☺	☺
1050	К	Л	М	Н	О	П	Р	С	Т	У	☺	☺	☺
1060	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Э	Э	☺	☺	☺
1070	Ю	Я	а	б	в	г	д	е	ж	з	☺	☺	☺
1080	и	й	к	л	м	н	о	п	р	с	☺	☺	☺
1090	т	у	ф	х	ц	ч	ш	щ	ъ	ы	☺	☺	☺
1100	ь	э	ю	я	è	ë	ђ	ѓ	є	ѕ	☺	☺	☺
1110	і	ї	ј	љ	њ	ќ	ќ	ў	љ	љ	☺	☺	☺
1120	ω	w	Ѡ	ѡ	Є	Ѕ	Ѕ	Ѕ	Ѕ	Ѕ	☺	☺	☺

Unicode (1988) dont UTF-8
 plus de 100 000 caractères
 (dont > 3 000 émojis)
 de 1 jusqu'à 6 octets

Des documents « images »

26866-26914

Royaume-Uni-United Kingdom

- 26866 Boss, Medard : A PSYCHIATRIST DISCOVERS INDIA. - Henry A. Frey. - London : Wolff, 1965. 192. 31/6. Dt: *Indienfahrt eines Psychiaters*
- 26867 Boutrais, Cyprien Burns & Oates. VII
- 26868 Buys, J. : CHRIST London : Chapmar du monde
- 26869 Carrier, Hervé : T - Arthur J. Arriery. 1965. 335. 30/- . *religieuse*
- 26870 Congar, Yves : I Loretz. - London *Chrétiens en dialogue*
- 26871 JESUS CHRIST. - 223. 25/- . Fr: *Jésus*
- 26872 TRADITION ANI Thomas Rainborou 90/- . Fr: *La traditi*
- 26873 Cristiani, Léon : London : Burns & croire
- 26874 Daniel-Rops, Hen 1870-1939. - John 48/- . Fr: *L'Eglise a*
- 26875 Desbuquois, Gust bury Wells : Fowle
- 26876 Dournes, Jacques : - London : G. Chapman. 205. 50/- . Fr: *Dieu aime les païens*
- 26877 Drze, A. : LIVING IN CHRIST, liturgy and sacraments. - F.M. Gale & Jennifer Nicholas. - London : G. Chapman. 209. 8/6. Fr: *Jésus Christ*
- 26878 Ebeling, Gerhard : THE L - James W. Leitch. - Lo Dt: *Vom Gebet : Predigten*
- 26879 THE NATURE OF FAITH. Collins. 191. 7/6. Dt: *Glaubers*
- 26880 THEOLOGY AND PROCLAI don : Collins. 186. 28/- . I
- 26881 ECUMENICAL DIALOGUE IN London : Lutterworth. 83. *les rencontres œcuméniques d*
- 26882 Emery, Pierre Y. : THE c M. Watson. - London : F: *croissants au ciel et sur la t*
- 26883 Floristan, Casiano : THE P John F. Byrne. - London 11/6. Esp: *La parroquia, e*
- 26884 Geissmann, Josef R. : W.J. O'Hara. - London : *Die Heilige Schrift und die*
- 26885 Häring, Bernhard : THE N London : Burns & Oates, I *Auftrag der Sakramente*
- 26886 Héris, Charles V. : SPIRIT - London : Herder, 1965. *l'amour*
- 26887 Jaeger, Lorenz *Erzbischof* ECUMENISM : THE COUNCI London : G. Chapman, I *Konzildekret über den Ökum*
- 26888 Jeanne d'Arc, *Seur* : V Martin Murphy. - London *Les religieuses dans l'Eglise*
- 26889 Jungmann, Josef A. : THI L. Batley. - London : Bu... *Das Eucharistische Hochgebet*
- 26890 LITURGICAL RENEWAL IN RETROSPECT AND PROSPECT. - Clifford Howell. - London : Burns & Oates, 1965. 45. 4/6. Dt: *Liturgische Erneuerung*
- 26891 THE LITURGY OF THE WORD. - H.E. Winstone. - London : Burns & Oates. 82. 8/6. Dt: *Wortgottesdienst im Lichte von*
- 26901 LITURGY IN DEVELOPMENT. - H.J.J. Vaughan. - London : Sheed & Ward, 1965. IX, 187, 12/6. Ned: *Medien in actualiteit*
- 26914 Roman Catholic Church. *2d Vatican Ecumenical Council.* DECLARATION ON RELIGIOUS LIBERTY. - Thomas Athillat. - London : Catholic Truth Society. 19. 1/- . Lat: *De libertate religiosa*

don : Collins. 186, 28/- . Dt: *Theologie und Verkündigung*

26881 ECUMENICAL DIALOGUE IN EUROPE. - Fletcher Fleet. - London : Lutterworth. 83, 12/6. Dt: *Dialogue œcuménique, les rencontres œcuméniques des Dombes*

26882 Emery, Pierre Y. : THE COMMUNION OF SAINTS. - D.J. & M. Watson. - London : Faith P. XIII, 256. Fr: *L'unité des*

2300 don : Collins. 186, 28/. Dt: Theologie und Verkündigung

26881 ECUMENICAL DIALOGUE IN EUROPE. - Fletcher Fleet. - London : Lutterworth. 83, 12/6. Dt: Dialogue œcuménique, les rencontres œcuméniques des Dombes

26882 Emery, Pierre Y. : THE COMMUNION OF SAINTS. - D.J. & M. Watson. - London : Faith P. XIII, 256. Fr: L'unité des

gallica.bnf.fr

BnF Gallica TOUT GALLICA Rechercher... RECHERCHE AVANCÉE

TOUTES NOS SÉLECTIONS PAR TYPES DE DOCUMENTS PAR THÉMATIQUES PAR AIRES GÉOGRAPHIQUES BLOG

Accueil > Consultation

Les fleurs du mal / par Charles Baudelaire

Baudelaire, Charles (1821-1867). Auteur du texte

SYNTHÈSE

EN SAVOIR PLUS

LÉGENDES ET TABLE DES MATIÈRES

VERSION TEXTE (OCR)

.MJ LECTEUR

La sottise, l'erreur, le péché, la lésine,
Occupent nos esprits et travaillent nos corps,
Et nous alimentons nos aimables remords,
Comme les mendiants nourrissent leur vermine.
Nos péchés sont têtus, nos repentirs sont lâches;
Nous nous faisons payer grassement nos aveux,
Et nous rentrons gaiement dans le chemin bourbeux
Croyant par de vils pleurs laver toutes nos taches.

i

1

Le taux de reconnaissance estimé pour ce document est de **86.86%**.
En savoir plus sur l'OCR

. AU LECTEUR

La sottise, l'erreur, le péché, la lésine,
Occupent nos esprits et travaillent nos corps,
Et nous alimentons nos aimables remords,
Comme les mendiants nourrissent leur vermine.

Nos péchés sont têtus, nos repentirs sont lâches;
Nous nous faisons payer grassement nos aveux,
Et nous rentrons gaiement dans le chemin bourbeux,
Croyant par de vils pleurs laver toutes nos taches.

1

A DÉCOUVRIR

ZOOM Page 1



L'encodage des documents : formats ouverts et méta-données



```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ns1="http://standoff.proposal"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="https://xml-schema.delivery.istex.fr/formats/tei-istex.xsd">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a system for
          design</title>
      </titleStmt>
      <publicationStmt>
        <authority>ISTEX</authority>
        <publisher ref="https://scientific-publisher.data.istex.fr/ark:/67375/H02-SWLMH5L1-1">Spring
        <pubPlace>London</pubPlace>
        <availability>
          <licence>Springer-Verlag London Limited</licence>
          <p scheme="https://loaded-corpus.data.istex.fr/ark:/67375/XBH-3XSW68JL-F">springer</p>
        </availability>
        <date type="published" when="2006">2006</date>
      </publicationStmt>
      <notesStmt>
        <note type="content-type"
          subtype="research-article"
          source="OriginalPaper"
          scheme="https://content-type.data.istex.fr/ark:/67375/XTP-1JC4F85T-7">research-article
        <note type="publication-type"
          subtype="journal"
          scheme="https://publication-type.data.istex.fr/ark:/67375/JMC-5WTPMB5N-F">journal</not
      </notesStmt>
      <sourceDesc>
        <biblStruct>
          <analytic>
            <title level="a" type="main" xml:lang="en">The design and realization of CoViD: a syst
              design</title>
            <author role="corresp">
              <persName>
                <forename type="first">Wolfgang</forename>
                <surname>Stuerzlinger</surname>
              </persName>
              <affiliation>
                <orgName type="institution">York University</orgName>
                <address>
                  <settlement>Toronto</settlement>
                  <country key="CA" xml:lang="en">CANADA</country>
                </address>
              </affiliation>
            </analytic>
          </biblStruct>
        </sourceDesc>
      </teiHeader>
```

XML

Abstract Many important decisions in the design process are made during fairly early on, after designers have presented initial concepts. In many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today's systems and tools do not support this, as computer systems typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a new system to facilitate a collaborative 3D design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design systems. Then, we introduce the new system, which consists of two main parts. The first part is an easy-to-use conceptual 3D design tool that can be used productively even by naive users. The tool provides novel interaction techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and modification of 3D models. The second part is a novel infrastructure for collaborative work, which of a semi-immersive

W. Stuerzlinger (&&) L. Zaman A. Pavlovych
York University, Toronto, Canada
URL: <http://www.cs.yorku.ca/~wolfgang>
URL: <http://www.cs.yorku.ca/~zaman>
URL: <http://www.cs.yorku.ca/~andriyp>
J.-Y. Oh
University of Arizona, Tucson, AZ, USA
e-mail: jyoh@optics.arizona.edu

setup. It is designed to support multiple users working together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present against the guidelines for collaborative 3D design. Finally, we present re which asked naive users to collaborate in a 3D design task on the new system.
Keywords Collaborative design 3D design
Collaborative virtual reality

<https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>

L'encodage des documents : formats ouverts et méta-données

OpenEdition Books

ACCUEIL CATALOGUE DES 10962 LIVRES ÉDITEURS AUTEURS OPENEDITION S

LA FRANCE D'ANTONIO GRAMSCI | Romain Descendre, Jean-Claude Zancarini

Gramsci et les Lumières

Giuseppe Cospito

p. 39-58

RECHERCHER DANS LE LIVRE

TABLE DES MATIÈRES

CITER PARTAGER

AJOUTER À ORCID

Observations préliminaires

1 L'objectif des pages qui vont suivre est...

```
<link rel="Contents" href="http://books.openedition.org/enseditions/17004"/>
<link rel="Start" href="http://books.openedition.org/enseditions/17054" title="Remerciements"/>
<link rel="Prev" href="http://books.openedition.org/enseditions/17064" title="La France de Gramsci"/>
<link rel="Next" href="http://books.openedition.org/enseditions/17079" title="Gramsci et Rousseau"/>
<meta name="DC.title" content="Gramsci et les Lumières"/>
<meta name="description" xml:lang="fr" lang="fr" content="Observations préliminaires L'objectif des pages qui vont s
ouve dans les écrits de Gramsci, en partant des premiers articles de journaux jusqu'aux écrits de prison. Afin d'évite
s figures particulières ou des courants du xviiiè siècle. No..."/>
<meta name="thumbnail" content="https://static.openedition.org/covers/OB/enseditions/17004/17004-120x240.jpg"/>
<meta name="copyright" content="© ENS Éditions, 2021 Conditions d'utilisation : http://www.openedition.org/6540"/>
<meta name="DC.rights" content="© ENS Éditions, 2021 Conditions d'utilisation : http://www.openedition.org/6540"/>
<meta name="author" content="Cospito, Giuseppe"/>
<meta name="DC.creator" content="Cospito, Giuseppe"/>
<meta name="DC.publisher" content="ENS Éditions"/>
<meta name="DC.date" scheme="W3CDTF" content="2021"/>
<meta name="DC.identifiant" scheme="ISBN" content="9791036202735"/>
```

15 Gramsci décrira dans les *Cahiers* ce moment comme un passage de la phase « économique-corporative » à la phase proprement « politique » de la lutte pour l'émancipation d'un groupe social. Selon lui,

le dernier exemple, le plus proche de nous, et par conséquent le moins différent de notre cas est celui de la Révolution française. La période culturelle antérieure, dite des Lumières, si décriée par les critiques superficielles de la raison théorique, ne fut pas du tout ou du moins ne se limita pas à être ce papillonnement de beaux esprits encyclopédiques qui discouraient de tout et de tous avec une égale imperturbabilité, et croyaient n'être hommes de leur temps qu'après avoir lu la grande *Encyclopédie* de d'Alembert et de Diderot. En somme, ce ne fut pas seulement un phénomène d'intellectualisme pédant et aride, pareil à celui que nous avons sous les yeux qui trouve son déploiement maximum dans les Universités populaires de dernier ordre. En soi, ce fut une magnifique révolution par laquelle, comme le remarque pertinemment De Sanctis dans son *Histoire de la littérature italienne*, s'était formée dans toute l'Europe une sorte de conscience unitaire, une internationale spirituelle bourgeoise, sensible en chacun de ses éléments aux douleurs et aux malheurs communs, et qui constituait la meilleure des préparations à la révolte sanglante qui se réalisa ensuite en France. [Grâce à cela,] les baïonnettes des armées de Napoléon trouvaient la voie.

HTML5

ePUB

```
<?xml version="1.0"
<!DOCTYPE html PUBLIC
<html xmlns="http://www.w3
<head><title>XYZ</title>
</head>
<body>
<p>
voluptatem accusantium do
totam rem aperiam eaque
</p>
</body>
</html>
```



OpenEdition est un portail de ressources électroniques en sciences humaines et sociales. Si vous souhaitez que votre établissement s'abonne à des services complémentaires et vous donne accès à des formats détachables (PDF, ePub), consultez les pages Institutions.



PLUS D'INFORMATIONS

OpenEdition Books

À la Une

Nouveaux dossiers

Nouveaux extraits

arXiv is a free distribution service and an open-access archive for 1,826,638 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse:

Physics Search Form interface Catchup

News

Read about recent news and updates on arXiv's blog. (View the former "what's new" page here). Read about how to prevent attempting any automa

Physics

- Astrophysics (astro) includes: Astrophysics, Astrophysical Phenomena
- Condensed Matter includes: Disordered Systems, Quantum Gases, Soft Matter
- General Relativity and Quantum Gravity
- High Energy Physics

COVID-19 Quick Links

See COVID-19 SARS-CoV-2 preprints from

- arXiv
- medRxiv and bioRxiv

Important: e-prints posted on arXiv are not peer-reviewed by arXiv; they should not be relied upon without context to guide clinical practice or health-related behavior and should not be reported in news media as established information without consulting multiple experts in the field.

ISTEX

23 millions de documents provenant de 30 corpus de littérature scientifique dans toutes les disciplines, soit plus de 9 314 revues et 348 636 ebooks entre 1473 et 2019 pour l'ESR

Testez ISTEX : indiquez un titre, des mots-clés ou un DOI



Vous êtes ?



CHERCHEUR

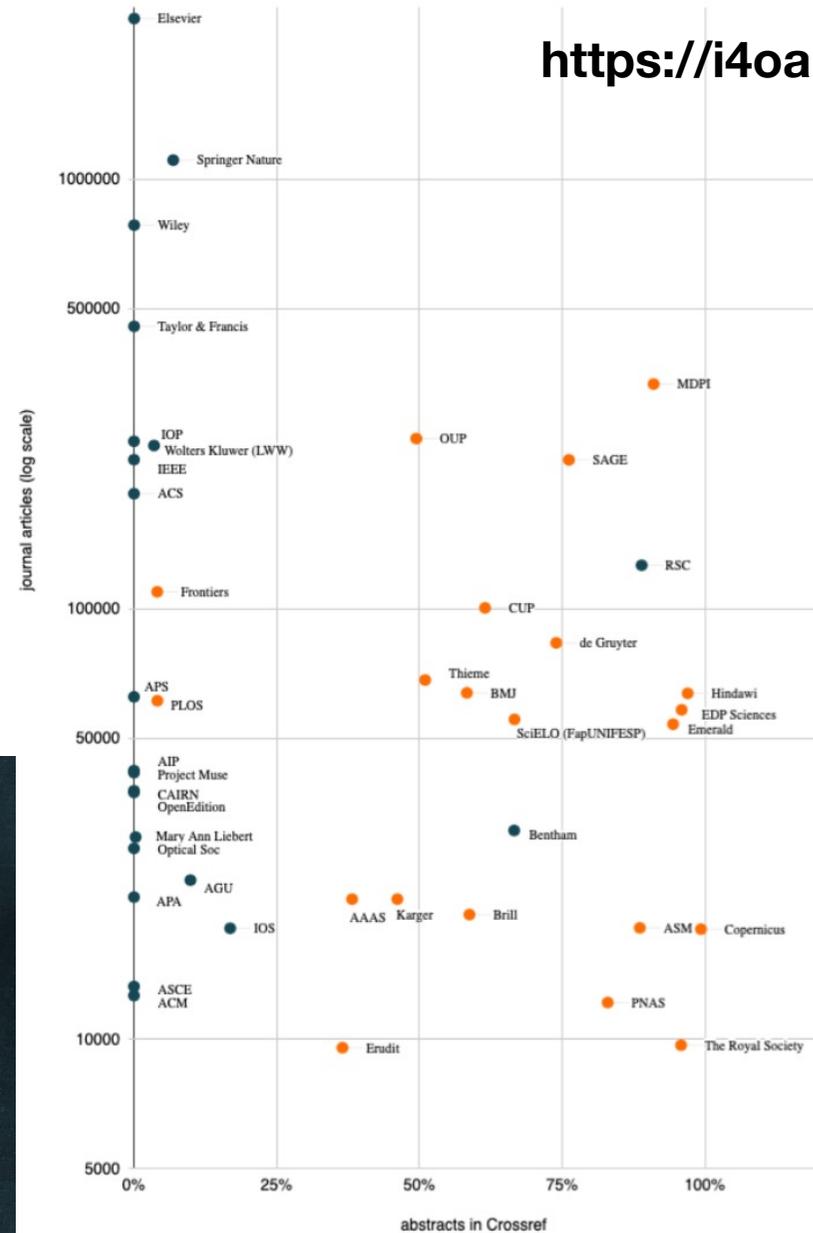


RESPONSABLE



CURIEUX

selected publishers - abstracts in Crossref journal articles (2018-2020) per 2021-01-01



<https://i4oa.org>

D'autres formats ouverts

```
"jour";"nomReg";"numReg";"incid_rea"  
2020-03-19;"Auvergne-Rhône-Alpes";84;44  
2020-03-19;"Bourgogne-Franche-Comté";27;33  
2020-03-19;"Bretagne";53;8  
2020-03-19;"Centre-Val de Loire";24;6  
2020-03-19;"Corse";94;11  
2020-03-19;"Grand-Est";44;69  
2020-03-19;"Guadeloupe";1;0  
2020-03-19;"Guyane";3;0  
2020-03-19;"Hauts-de-France";32;37  
2020-03-19;"Ile-de-France";11;151  
2020-03-19;"La Réunion";4;0  
2020-03-19;"Martinique";2;0  
2020-03-19;"Mayotte";6;0  
2020-03-19;"Normandie";28;7  
2020-03-19;"Nouvelle-Aquitaine";75;7  
2020-03-19;"Occitanie";76;29  
2020-03-19;"Pays de la Loire";52;11  
2020-03-19;"Provence-Alpes-Côte d'Azur";93;25  
2020-03-20;"Auvergne-Rhône-Alpes";84;16  
2020-03-20;"Bourgogne-Franche-Comté";27;9  
2020-03-20;"Bretagne";53;2  
2020-03-20;"Centre-Val de Loire";24;4  
2020-03-20;"Corse";94;0  
2020-03-20;"Grand-Est";44;45
```

CSV

https://en.wikipedia.org/wiki/Comma-separated_values

```
"header": {  
  "title": "The JSON example",  
  "descriptionText": "This is some title text."  
},  
"content": {  
  "title": "The content example text",  
  "elements": [  
    {  
      "title": "The first element",  
      "mainText": "First element main text",  
      "additionalText": "First element additional te  
    },  
    {  
      "title": "The second element",  
      "mainText": "Second element main text",  
      "additionalText": "Second element additional
```

JSON

<https://en.wikipedia.org/wiki/JSON>

```
---  
receipt:      Oz-Ware Purchase Invoice  
date:        2012-08-06  
customer:  
  first_name: Dorothy  
  family_name: Gale  
  
items:  
- part_no:    A4786  
  descrip:    Water Bucket (Filled)  
  price:      1.47  
  quantity:   4  
  
- part_no:    E1628  
  descrip:    High Heeled "Ruby" Slippers  
  size:       8  
  price:      133.7  
  quantity:   1  
  
bill-to: &id001  
  street: |  
          123 Tornado Alley  
          Suite 16  
  city:     East Centerville  
  state:    KS  
  
ship-to: *id001  
  
specialDelivery: >  
  Follow the Yellow Brick
```

YAML

<https://en.wikipedia.org/wiki/YAML>

Niveaux lexicaux, syntaxiques, sémantiques

Stanford CoreNLP 4.2.0 (updated 2020-11-16)

<https://corenlp.run>

— Text to annotate —

Parfois dans ces derniers jours d'hiver, nous entrions avant d'aller nous promener dans quelqu'une des petites expositions qui s'ouvraient alors et où Swann, collectionneur de marque, était salué avec une particulière déférence par les marchands de tableaux chez qui elles avaient lieu.

— Annotations —

parts-of-speech x constituency parse x dependency parse x

— Language —

French

Submit

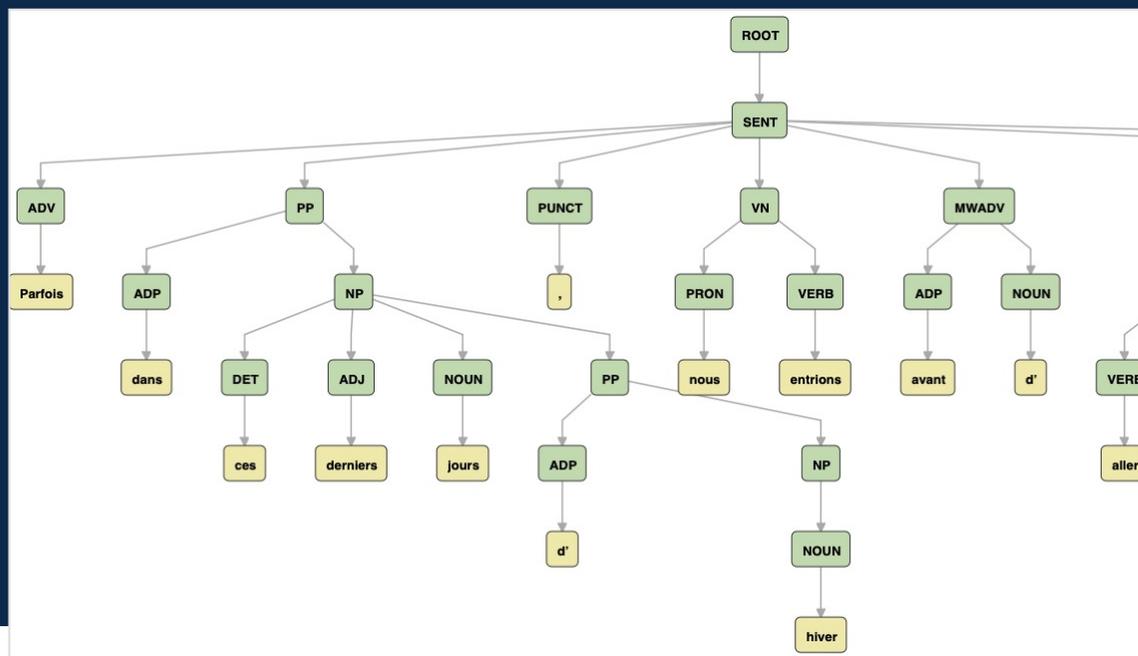
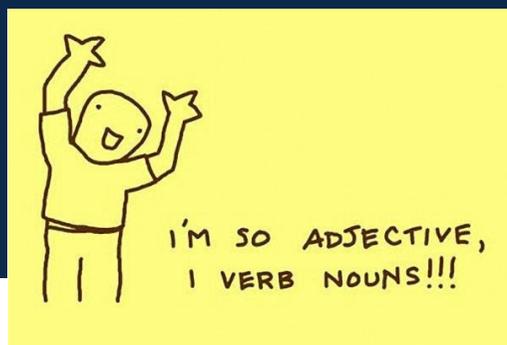
Part-of-Speech:

1 Parfois dans ces derniers jours d' hiver, nous entrions avant d' aller nous promener dans quelqu'une de les petites expositions qui s' ouvraient alors et où Swann, collectionneur de marque, était salué avec une particulière déférence par les marchands de tableaux chez qui elles avaient lieu.

Dépendances

Unités lexicales et parties du discours

Constituants



? Answer a question ^

Reading Comprehension

Visual Question Answering

< Annotate a sentence ^

Named Entity Recognition

Open Information Extraction

Sentiment Analysis

Dependency Parsing

Constituency Parsing

Semantic Role Labeling

≡ Annotate a passage ^

Coreference Resolution

✂ Generate a passage ^

Language Modeling

Masked Language Modeling

⚖ Compare two sentences ^

Textual Entailment

Named Entity Recognition

Named Entity Recognition is the task of identifying named entities (people, locations, organizations, etc.) in the input text.

Model

Fine Grained Named Entity Recognition

This model identifies a broad range of 16 semantic types in the input text. It is a reimplementation of Lample (2016) and uses a biLSTM with a CRF layer, character embeddings and ELMo embeddings.

[TaskDemo](#)

[Model Card](#)

[Model Usage](#)

Example Inputs

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there.

Sentence

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there.

Run Model

Model Output

Share

Entities

When I told John that I wanted to move to Alaska, he warned me that I'd have trouble finding a Starbucks there .

PERSON GPE ORG

Enrichir par la création de liens entre documents



The screenshot shows a web browser window titled "Bilbo - web" with the URL "bilbo.openeditionlab.org". The main content area displays XML code for a bibliography list. On the right side, there are controls for "Corpus 1 (bibliography)" and "Corpus 2 (notes)", including "Annotate" and "Reset" buttons, and "Test corpus 1" and "Test corpus 2" buttons. Below the XML code, a list of references is displayed, including:

- Bortoli M., Cutini V., 1999, Accessibilità urbana e distribuzione delle attività. L'analisi configurazionale del centro storico di Volterra, in Atti della XX Conferenza Italiana di Scienze Regionali, Piacenza, 5-7 Ottobre.*
- Hill D.M., Bakker J.J., Akers B.L., 1964, An Evaluation of the Needs of the Pedestrian in Downtown, Traffic Research Corporation, Chicago.*
- Hillier B., 1996, Space is the Machine, Cambridge University Press, Cambridge.*
- Hillier B., 1999, Why space syntax works, when it looks as though it should not, in Environment & Planning B : Planning and Design, numero speciale monografico sullo Space Syntax Symposium (in corso di pubblicazione).*
- Hillier B., Hanson J., 1984, The Social Logic of Space, Cambridge University Press, Cambridge.*
- Hillier B., Penn A., Hanson J., Grajevski, Xu J., 1993, Natural Movement : or, Configuration and Attraction in Urban Pedestrian Movement, in Enviroment & Planning B, Planning and Design, vol. 20.*
- Hoel L.A., 1968, Pedestrian Travel Rates in Central Business Districts, in Traffic Engineering and Control, January, 10-13.*
- Lautso K., Murola P., 1974, A Study of Pedestrian Traffic in Helsinki, in Traffic Engineering and Control, January, 446-449.*
- O'Flaherty C.A., Parkinson M.H., 1972, Movement on a City Centre Footway, in Traffic Engineering and Control, February, 434-438.*
- Pushkarev B., Zupan J., 1975, Urban Space for Pedestrians, MIT Press, Cambridge, MA.*

At the bottom of the page, there is a footer with the text: "spatial simulation », *Computers, Environment and Urban Systems*, vol. 32, No.6, 417-430. DOI : 10.1016/j.compenvurbsys.2008.09.004"

Test : <http://bilbo.openeditionlab.org>

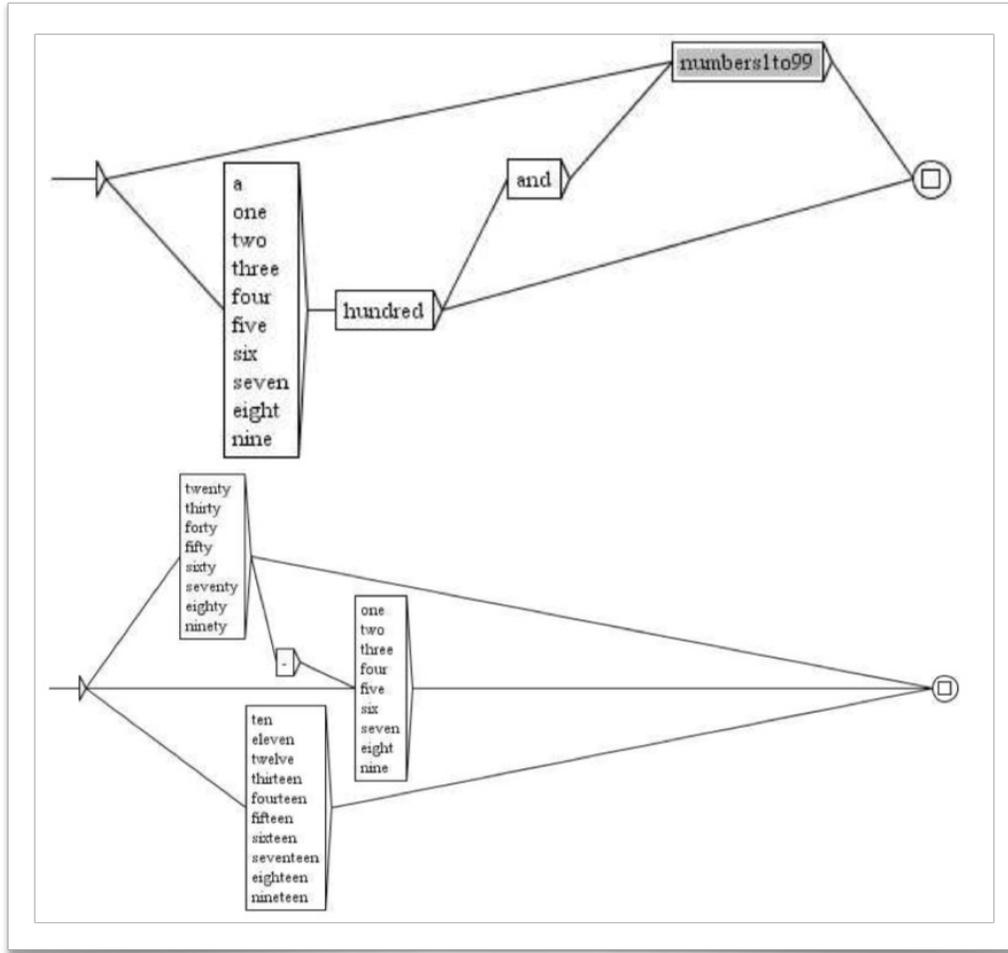
Sources : <http://github.com/OpenEdition/bilbo>



Des approches informatiques et des ressources linguistiques

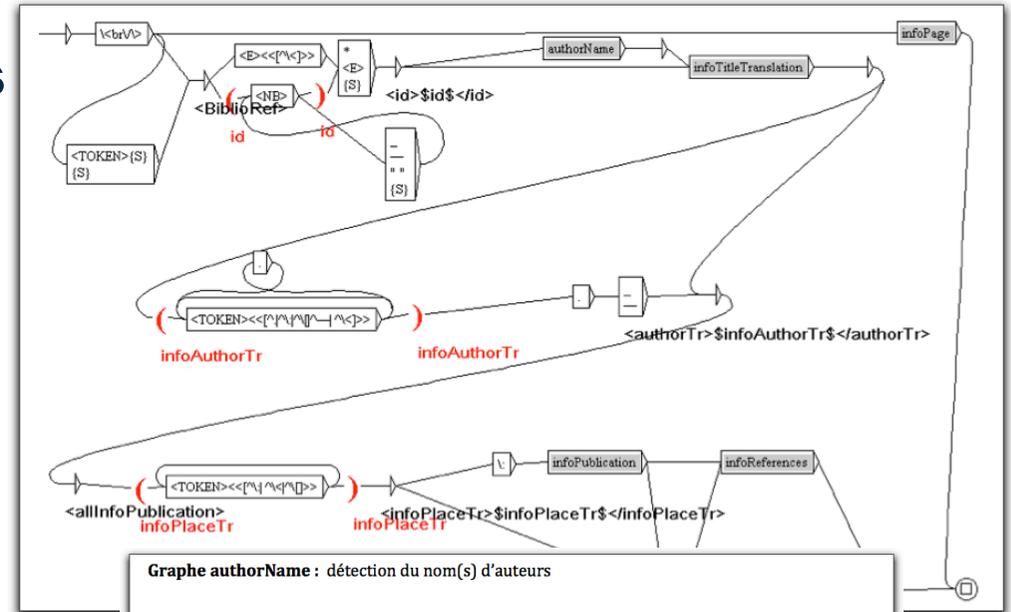
04.06.19

Des exemples d'automates

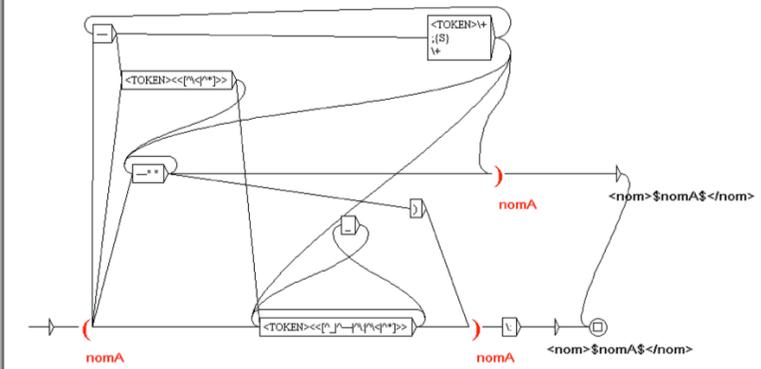


Eric Laporte. Symbolic Natural Language Processing. Lothaire. Applied Combinatorics on Words, Cambridge University Press, pp.164-209, 2005. hal-00145253

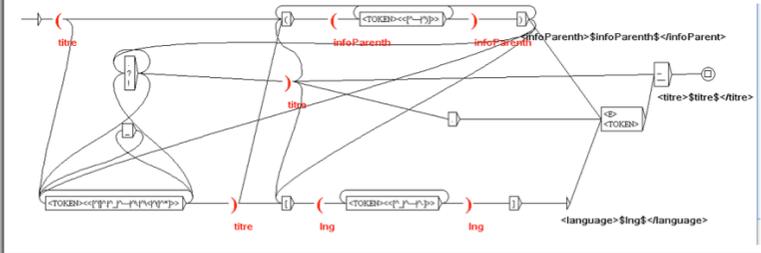
<https://hal.archives-ouvertes.fr/hal-00145253/document>



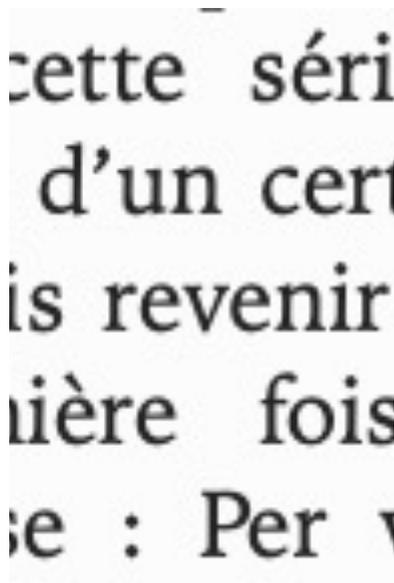
Grappe authorName : détection du nom(s) d'auteurs



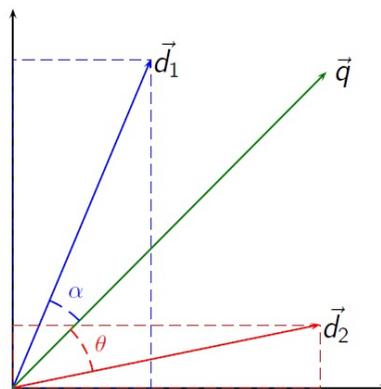
Grappe infoTitleTranslation : détection du titre de la traduction



Des formes, des distributions et des vecteurs



Des formes rassemblées, des sacs (de lettres, de mots)



Des vecteurs

Mot	Probabilité

Des modèles de langues

$$P(\text{Classe} | (w_1, w_2, \dots, w_{T-1}, w_T))$$

$$\text{Similarité}(d_1, d_2) \approx \vec{d}_1 \cdot \vec{d}_2$$

$$\text{Similarité}(d_1, d_2) \approx \cos(\vec{d}_1, \vec{d}_2)$$

$$P(w_1, w_2, \dots, w_{T-1}, w_T) = \prod_{t=1}^T P(w_t | w_{t-1}, w_{t-2}, \dots, w_1)$$

the	cat	sat	on	the	mat	$P(w_1)$
the	cat	sat	on	the	mat	$P(w_2 w_1)$
the	cat	sat	on	the	mat	$P(w_3 w_2, w_1)$
the	cat	sat	on	the	mat	$P(w_4 w_3, w_2, w_1)$
the	cat	sat	on	the	mat	$P(w_5 w_4, w_3, w_2, w_1)$
the	cat	sat	on	the	mat	$P(w_6 w_5, w_4, w_3, w_2, w_1)$

Slide Credit: Piotr Mirowski

Un exemple de modèle de langue



Ngram Viewer Exports

The Google Books Ngram Viewer is optimized for quick inquiries into the usage of small sets of phrases. If you're interested prefer to download a portion of the corpora yourself. Or all of it, if you have the bandwidth and space. We're happy to oblige.

These datasets were generated in February 2020 (version 3), July 2012 (Version 2) and July 2009 (Version 1); we will update versions will have distinct and persistent version identifiers (20200217, 20120701 and 20090715 for the current sets).

Each of the numbered links below will directly download a fragment of the corpus. In Version 2 the ngrams are grouped alpha Version 1 the ngrams are partitioned into files of equal size. In addition, for each corpus we provide a file named `total_count` that make up the corpus. This file is useful for computing the relative frequencies of ngrams.

A summary of how the corpora were constructed can be found [here](#). We explain it in greater depth [here](#) (Version 2) and [here](#) appear over 40 times across the corpus. That's why the sum of the 1-gram occurrences in any given corpus is smaller than th

File format: Each of the files below is compressed *tab*-separated data. In Version 2 each line has the following format:

```
ngram TAB year TAB match_count TAB volume_count NEWLINE
```

Language	#Volumes	#Tokens
English	4,541,627	468,491,999,592
Spanish	854,649	83,967,471,303
French	792,118	102,174,681,393
German	657,991	64,784,628,286
Russian	591,310	67,137,666,353
Italian	305,763	40,288,810,817
Chinese	302,652	26,859,461,025
Hebrew	70,636	8,172,543,728

Table 1: Number of volumes and tokens for each language in our corpus. The total collection contains more than 6% of all books ever published.

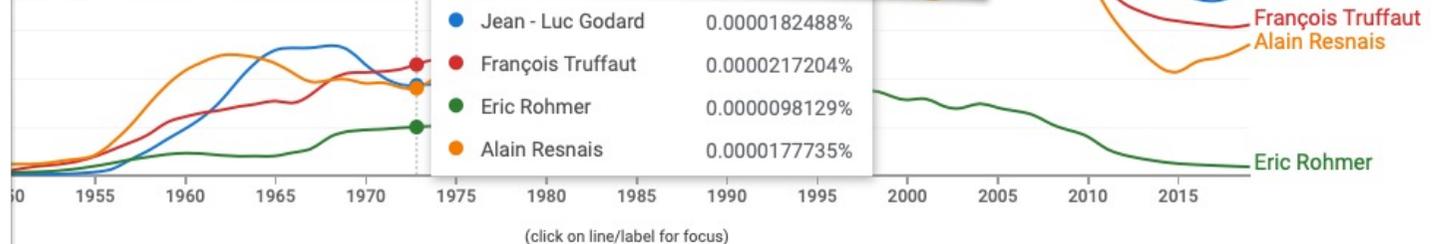
French

Version 20200217

- [1-grams](#)
- [2-grams](#)
- [3-grams](#)
- [4-grams](#)
- [5-grams](#)
- [Dependencies](#)

Version 20120701

[total_counts](#)



<https://books.google.com/ngrams>

[Reading Comprehension](#)[Visual Question Answering](#)[Annotate a sentence](#)[Named Entity Recognition](#)[Open Information Extraction](#)[Sentiment Analysis](#)[Dependency Parsing](#)[Constituency Parsing](#)[Semantic Role Labeling](#)[Annotate a passage](#)[Coreference Resolution](#)[Generate a passage](#)[Language Modeling](#)[Masked Language Modeling](#)[Compare two sentences](#)[Textual Entailment](#)

Language Modeling

Language modeling is the task of determining the probability of a given sequence of words occurring in a sentence.

Model

GPT2-based Next Token Language Model

This is the public 345M parameter OpenAI GPT-2 language model for generating sentences. The model embeds some input tokens, contextualizes them, then predicts the next word, computing a loss against known target. If `BeamSearch` is given, this model will predict a sequence of next tokens.

[TaskDemo](#)[Model Card](#)

Example Inputs

Sentence

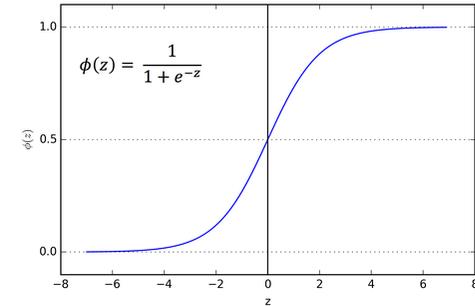
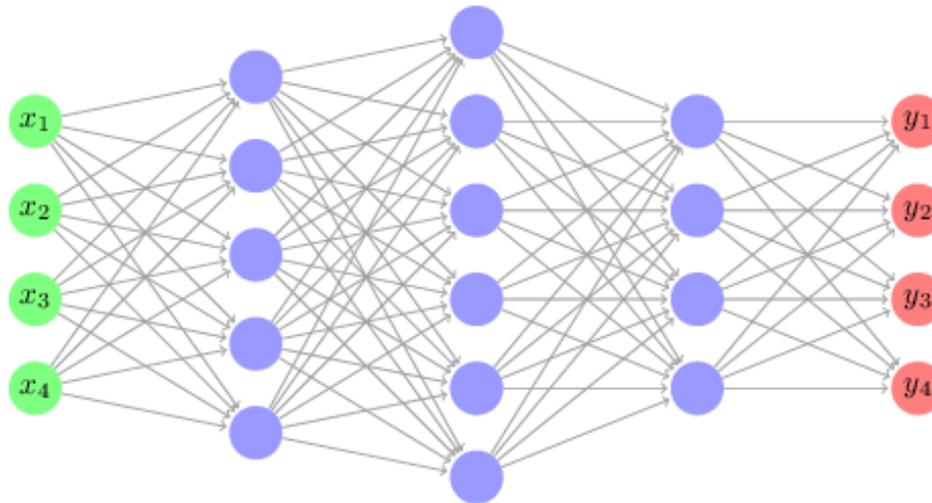
[Run Model](#)

Model Output

[Share](#)

Prediction	Score
The doctor ran to the emergency room to see the patient. ↵" ...	 99,1 %
The doctor ran to the emergency room to see the girl. She was crying ...	 0,6 %
The doctor ran to the emergency room to see the injured victim. ↵	 0,2 %

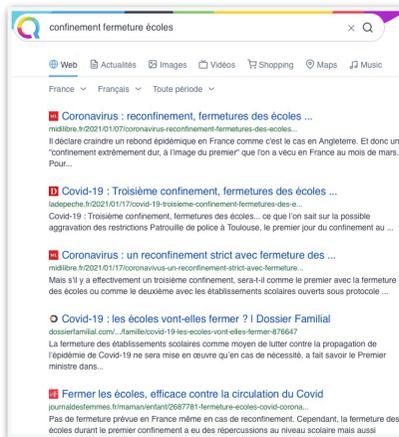
Approches numériques : des scores



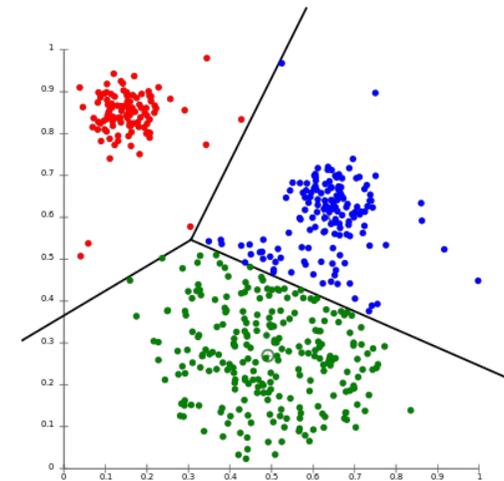
Data



Ranking / Scores



Décision ?



Home > Services > Virtual Language Observatory (VLO)

Virtual Language Observatory (VLO)



The Virtual Language Observatory (VLO) provides a means of exploring language resources and tools. Its aim is to provide an easy to use interface, allowing for a uniform search and discovery process for a large number of resources from a wide variety of domains. Facets make it easy to explore and access available resources. A powerful query syntax makes it possible to carry out more targeted searches as well. It also makes it easy to review processing options for discovered resources via the Language Resource Switchboard, and to create virtual collections based on search results via the Virtual Collection Registry.

[→ Go to the Virtual Language Observatory](#)

The following list provides a few links for example selections and queries to start exploring:

- [Resources for spoken French](#)
- [Corpora with Polish content](#)
- [All records from the Language Bank of Finland](#)
- Searching for a general term: "slovenian news sentiment"
- Searching for a specific record or set of records: "Hamburg MapTask Corpus"

More information is available in the [VLO's integrated help page](#).



<https://www.clarin.eu/content/virtual-language-observatory-vlo>

ORTOLANG

Accueil

Rechercher dans ORTOLANG

Corpus

Lexiques

Terminologies

Outils

Projets Intégrés

Actualités

Informations

Producteurs

Outils et Ressources pour un Traitement Optimisé de la LANGue

ORTOLANG est un équipement d'excellence validé dans le cadre des [investissements d'avenir](#). Son but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui :

- permette, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international;
- facilite l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels, en particulier vers les PME qui souvent ne peuvent pas se permettre de développer de telles ressources et outils de traitement de la langue compte tenu de leurs coûts de réalisation;
- valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics.

Huma-Num ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par **Huma-Num** (très grande infrastructure de recherche).

La charte d'ORTOLANG définit les modalités d'utilisation et de dépôt de ressources sur la plate-forme. Vous pouvez [consulter la charte](#) ou la télécharger ([fichier au format pdf](#))

 ORTOLANG bénéficie d'une aide de l'Etat au titre du programme « Investissements d'avenir » (ANR-11-EQPX-0032)
ORTOLANG ISSN 2417-7482

<https://www.ortolang.fr>

CHU ROUEN NORMANDIE

CiSMeF
Catalogue et Index des Sites Médicaux de langue Française

Projet CiSMeF

Tous les outils et services

Aide

Une réalisation du D2IM - CHU Hôpitaux de Rouen

S'inscrire



CiSMeF Catalogue et Index des Sites Médicaux de langue Française

J'aide CiSMeF

Recherche Doc'CiSMeF

Sélection de sites, articles et documents en libre accès

Pathologies, traitements, médicaments etc. **RECHERCHER**

tous les types

- uniquement les recommandations professionnelles
- uniquement les documents d'enseignement - Épreuves Classantes Nationales
- uniquement les documents grand public et les associations de patients
- uniquement les thèses et mémoires

Index alphabétique, Index thématique - Nouveautés : Quoi de neuf ? Alertes  

126 791 sites et documents le 02/12/2020

Informations COVID-19

HeTOP
Health Terminology - Ontologie Patient
Consulter le MeSH et les autres terminologies de santé

CRBM
Constructeur de Requêtes Bibliographiques Médicales
Interroger PubMed, CiSMeF, et LISSa en français

LISSa
Littérature Scientifique en Santé
Consulter la littérature médicale scientifique francophone
1 316 069 références le 02/12/2020

<http://www.chu-rouen.fr/cismef/>

openMIN7ED

SERVICES JOIN ABOUT NEWS CONTACT US

STEP 1

STEP 2

OpenMinTed
TDM SERVICES FOR ALL!

<http://openminted.eu/omtd-services/>

Search

Build

Sign In

Execute

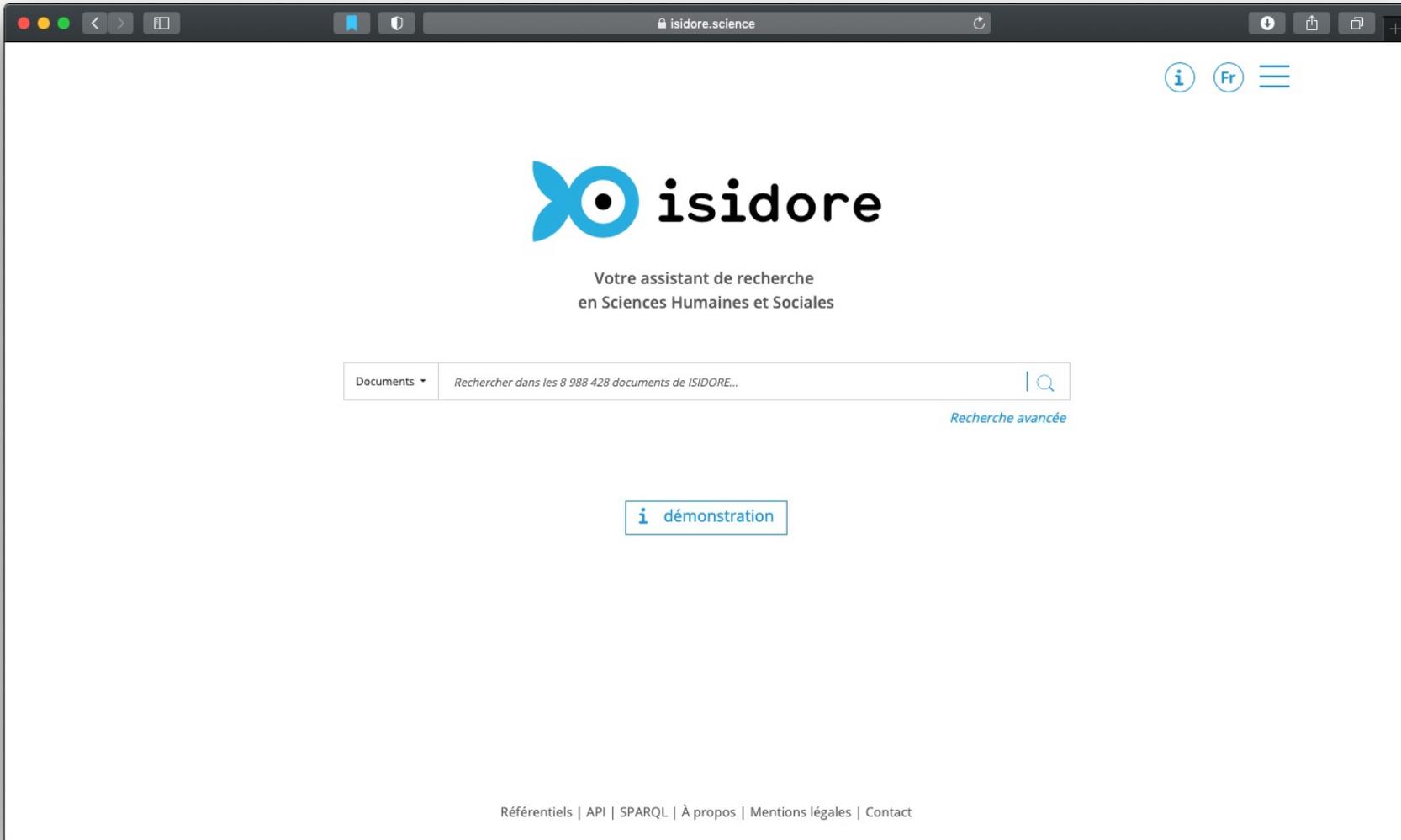
Annotate

<http://openminted.eu>



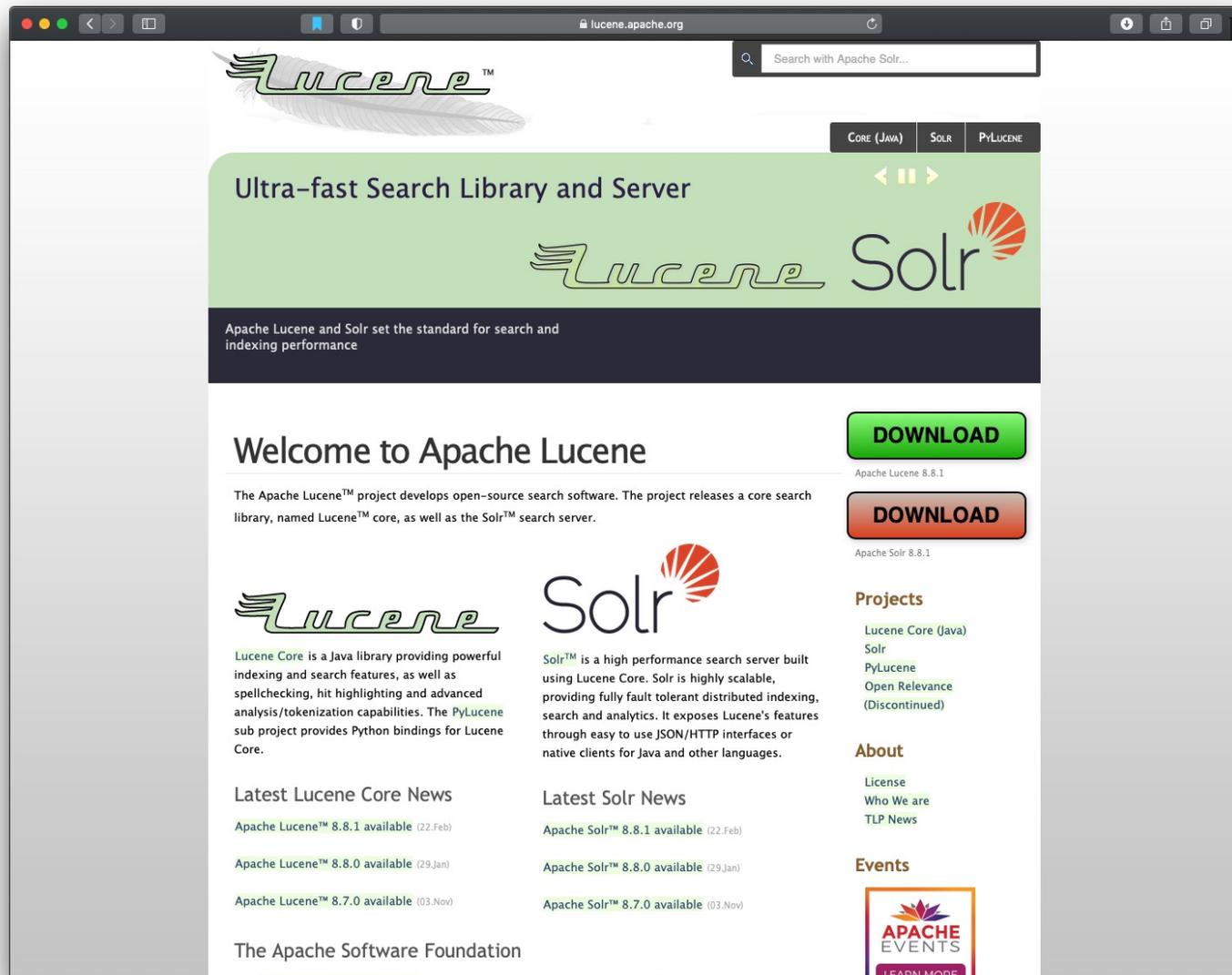
Des environnements logiciels pour le développement et l'expérimentation

Isidore : un moteur pour les SHS



<https://isidore.science>

Apache Lucene : indexation et recherche d'information



<https://lucene.apache.org>

Pour la recherche de données, x +

← → ↻ https://www.elastic.co/fr/products

elastic Produits Cloud Services Clients En savoir plus téléchargements contact Q FR

PRODUCTS

- Elasticsearch
- Kibana
- Logstash
- Beats
- ECE
- Features

SOLUTIONS

- Logging
- Metrics
- Site Search
- Security
- APM
- All



La Suite Elastic

La Suite Elastic, construite sur une fondation open source, vous permet de rechercher, analyser et visualiser, en toute fiabilité et sécurité, ainsi qu'en temps réel, des données issues de n'importe quelle source et sous n'importe quel format.

Ne manquez aucune info sur les mises à jour produit.

Adresse email Envoyer

En cliquant vous (1) acceptez [les conditions d'utilisation](#) et [la déclaration de confidentialité d'Elastic](#), et (2) acceptez de recevoir des e-mails occasionnels.



earthquake

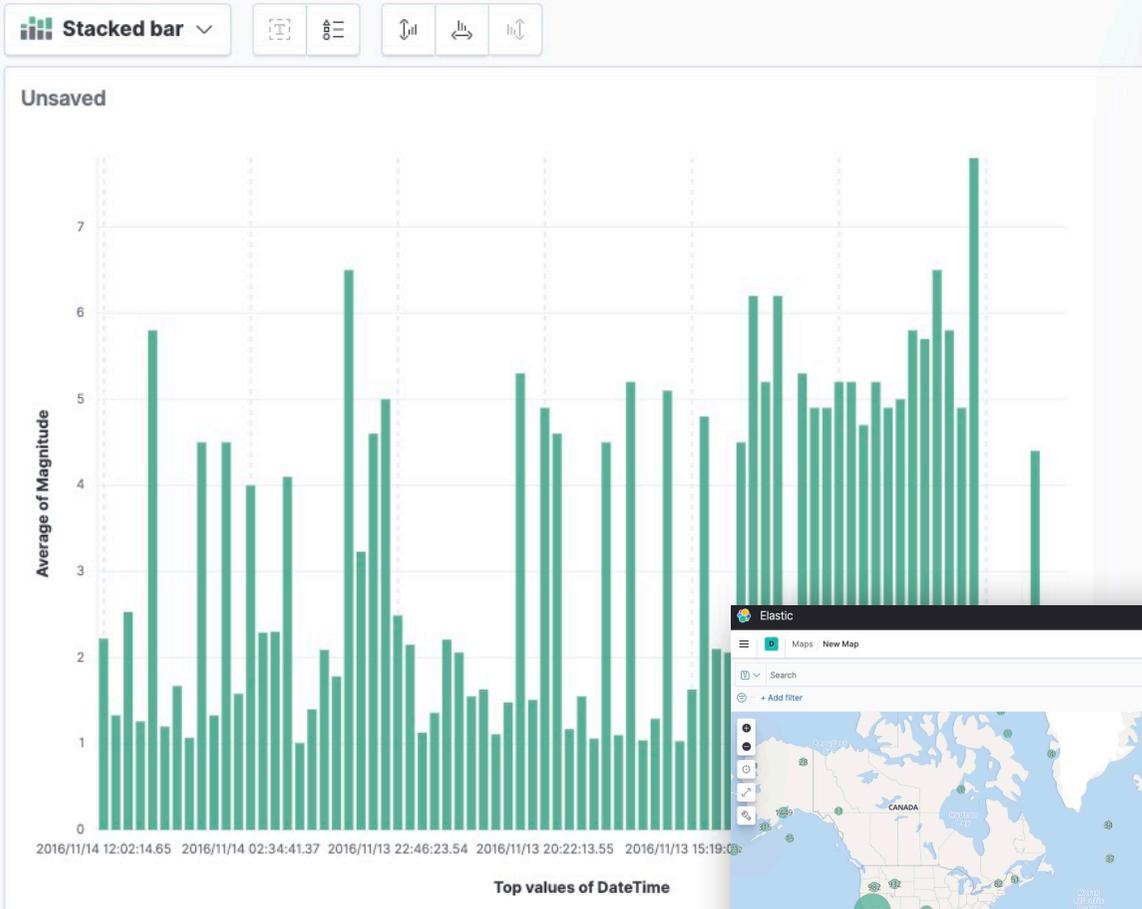
Search field names

Field filters 0

Records

Available fields 12

- DateTime
- Depth
- Distance
- EventID
- Gap
- Latitude
- Longitude
- Magnitude
- MagType
- NbStations
- RMS
- Source



X-axis configuration

Select a function

- Filters
- Intervals
- Top values

Select a field

DateTime

Number of values 1 to 100 (79)

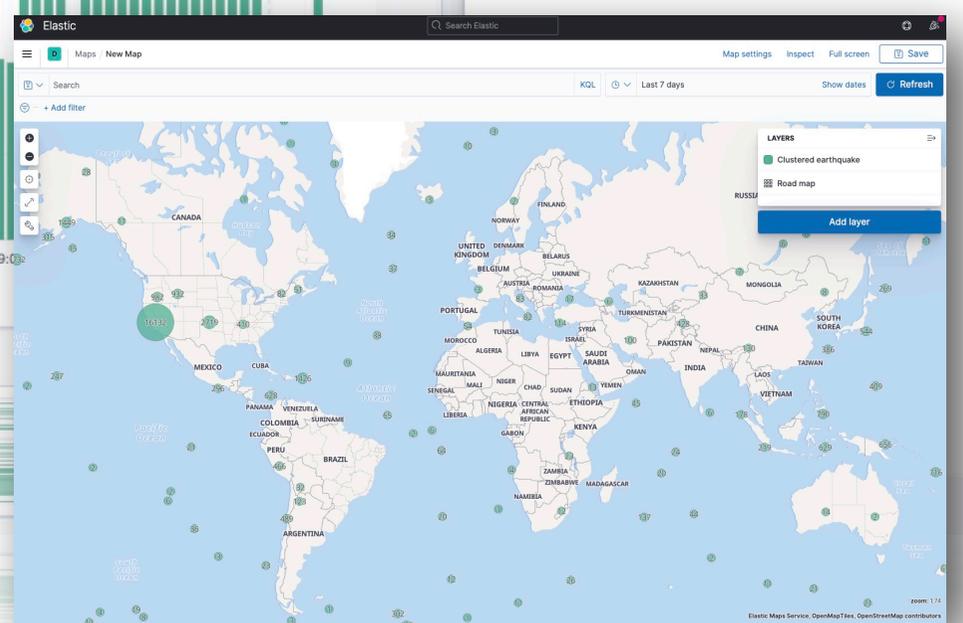
Order by Alphabetical

Order direction Descending

Display name Top values of DateTime

Suggestions

Current





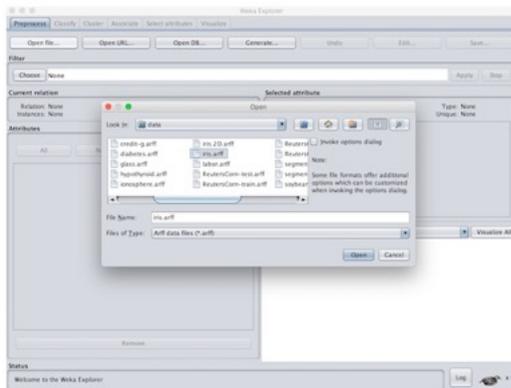
WEKA

The workbench for machine learning

Weka is a tried and tested open source machine learning workbench that can be accessed through a graphical user interface, terminal applications, or a Java API. It is widely used for

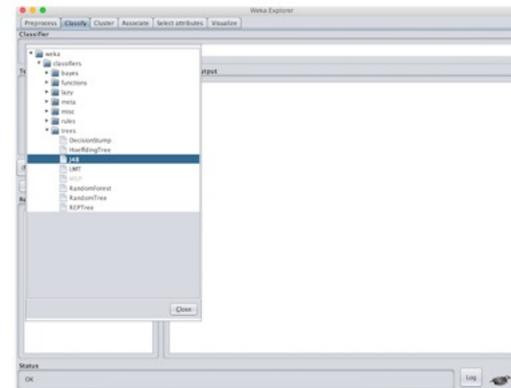
Machine Learning without Programming

Weka can be used to build machine learning pipelines, train classifiers, and run evaluations without having to write a single line of code:



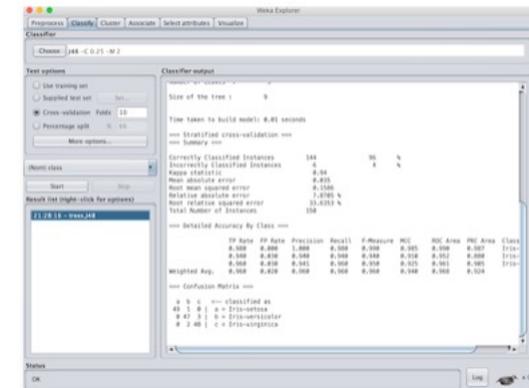
Open a dataset

First, we open the [dataset](#) that we would like to evaluate.



Choose a classifier

Second, we select a learning algorithm to use, e.g., the J48 classifier, which learns decision trees.



Evaluate predictive accuracy

Finally, we run a 10-fold cross-validation evaluation and obtain an estimate of predictive performance.

Unitex pour des approches symboliques

<http://igm.univ-mlv.fr/~unitex/>

Unitex/GramLab est une suite logicielle libre, multiplateforme, multilingue, fondée sur des dictionnaires et des grammaires pour l'analyse de corpus



Moteur
TAL

La **technologie fondée sur les automates** du moteur TAL de Unitex/GramLab permet de gérer des ressources électroniques telles que des dictionnaires et des grammaires électroniques et de les appliquer à un texte pour un traitement et une analyse rapides



Ressources
Linguistiques

Les ressources linguistiques sont des dictionnaires électroniques et des grammaires qui permettent l'analyse de données textuelles avec Unitex. Des ressources pour **plus de 22 langues** sont distribuées actuellement avec Unitex/GramLab



IDE
Visuel

L'Environnement de Développement Intégré visuel d'Unitex/GramLab permet aux utilisateurs de **concevoir et d'appliquer facilement des ressources linguistiques** aux textes. En outre, la déclinaison orientée projets permet d'exécuter des projets en un seul clic

règles syntaxiques ou sémantiques

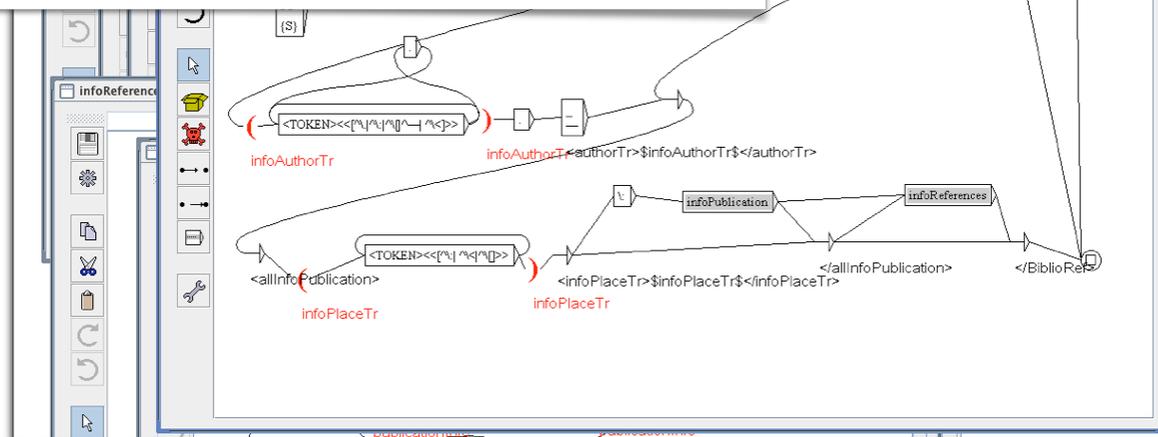
Analyse de Corpus

Construire, vérifier et appliquer des dictionnaires électroniques Appliquer les tables du lexique-grammaire

Aligner des textes Traiter l'ambiguïté grâce à l'automate du texte

Construire un automate à partir d'un corpus certifié

Rechercher des motifs à l'aide d'expressions régulières et de réseaux de transitions récursifs



- Overview ▸
- For Scientists ▸
- For Educators ▸
- For Business ▸
- News ▸

If you need to solve a problem with text analysis or language processing, you're in the right place.

- GATE is an **open source software toolkit** capable of solving almost any text processing problem.
- It has a **mature and extensive community** of developers, users, educators, students and researchers.
- It is used by **corporations, SMEs**, research labs and Universities worldwide.

File Options Tools Help

Annotations Sets Annotations List Annotations Stack Co-reference Editor OAT RATC RATI Text

Messages voting-example-... test-ontology-b...

2300 BST.

As well as picking MPs for Westminster, voters will elect councillors in 164 local authorities across England.

Voting in the general election will take place in 649 constituencies, with nearly 4,150 candidates standing for election across the country.

David Cameron was the first of the main UK party leaders to cast their vote. The Tory leader went to a community hall in Witney, Oxfordshire, shortly after 1030 BST, accompanied by his wife Samantha.

Labour leader Gordon Brown went to vote shortly after 1100 BST at a community centre close to his home in North Queensferry, Fife. His wife Sarah was with him.

Nick Clegg, leader of the Liberal Democrats, arrived at a polling station in Sheffield Hallam at 1120 BST. His wife Miriam is unable to vote in the general election because she

test-ontology-bdm-instar

- Entity
 - Location
 - Country
 - Organization
 - Person
 - Leader

Filter: [] X New Inst. Add to Selected Inst.

Instance	Label	Property	Value
Oxfordshire	[Oxfordshire]	partof	[Location]
Witney	[Witney]	partof	Oxfordshire

Document Editor Initialisation Parameters

The screenshot shows the Tag Editor interface with three sentences and their corresponding POS tags and dependency arcs. The TAG SET panel on the right lists various dependency types and POS tags.

Sentence 1: spaCy is fast and intuitive . '\n\n'

Sentence 2: It is a natural language processing library specifically

Sentence 3: spaCy is designed to help you do real work

TAG SET:

Dependencies	POS	NER
acl	acomp	amod
advcl	advmod	agent
appos	attr	aux
auxpass	case	cc
ccomp	compound	conj
cop	csubj	csubjpass
dative	dep	det
dobj	expl	intj
mark	meta	neg
nn	nmod	npadvmod
nsubj	nsubjpass	nummod
oprnd	obj	obl
parataxis	pcomp	pobj
poss	preconj	predet
prep	prt	punct
subtok	quantmod	relcl
ROOT	xcomp	

The screenshot shows the Tag Editor interface with four sentences and their corresponding sentiment and category tags. A 'Create DATA' dialog box is open, and the TAG SET panel on the right shows the selected categories.

Sentence 1: That was great!

Sentence 2: This is terrible.

Sentence 3: He likes to travel to Hawaii.

Sentence 4: The deal was a success, and everyone was happy.

Tags:

- Sentence 1: POSITIVE: True
- Sentence 2: POSITIVE: False
- Sentence 3: TRAVEL: True, POSITIVE: True
- Sentence 4: BUSINESS: True, POSITIVE: True

Create DATA Dialog:

- DATA options
- Include named entities
- Include dependencies
- Include POS tags
- Include CATS
- Add words
- Convert to JSON

TAG SET:

Dependencies	POS	CATS
Score:	True	False
POSITIVE	BUSINESS	TRAVEL
FASHION	FOOD	



prodigy

<https://prodi.gy/>

Radically efficient machine tea
An annotation tool powered
by active learning.

Named Entity Recognition

RTL CJK character-based

Try it live and highlight entities!

LOCATION 1 EVENT 2 DATE 3

وفي الفترة 1944-1945 افلح الجيش الكندي الأول في تحرير معظم أراضي هولندا LOCATION ، وكان يضم في صفوفه قوات كندية وبريطانية ويولندية . لكن سرعان ما بات لزاماً على الهولنديين بنهاية الحرب الأوروبية أن يحاربوا مقاتلي الثورة الوطنية الإندونيسية

SOURCE: ar.wikipedia.org/wiki/%D9%87%D9%88%D9%84%D9%86%D8%AF%D8%A7

Label any text, in any language or script

Prodigy lets you use token boundaries for faster and more consistent annotation, but it's also fully flexible: you can annotate from the character up if your task requires it. No matter what language or writing system you're working with, if it's text, Prodigy can help you annotate it.

Bootstrap with powerful patterns

Prodigy is a fully scriptable annotation tool, letting you **automate as much as possible** with custom rule-based logic. You don't want to waste time labeling every instance of common entities like "New York" or "the United States" by hand. Instead, give Prodigy rules or a list of examples, review the entities in context and annotate the exceptions. As you annotate, a statistical model can learn to suggest similar entities, generalising beyond your initial patterns.

patterns.json

```
{ "pattern": [ { "lower": "new" }, { "lower": "york" } ], "label": "CITY" }  
{ "pattern": [ { "lower": "berlin" } ], "label": "CITY" }
```

I live in **New York** CITY .



Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

GET STARTED

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

FACTS & FIGURES

Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

READ MORE

```
Edit the code & try spaCy spaCy v3.0 · Python 3 · via Binder

# pip install -U spacy
# python -m spacy download en_core_web_sm
import spacy

# Load English tokenizer, tagger, parser and NER
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("When Sebastian Thrun started working on self-driving cars at "
        "Google in 2007, few people outside of the company took him "
        "seriously. "I can tell you very senior CEOs of major American "
        "car companies would shake my hand and turn away because I wasn't "
        "worth talking to," said Thrun, in an interview with Recode earlier "
        "this week.")
doc = nlp(text)

# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)

RUN
```

```
Noun phrases: ['Sebastian Thrun', 'self-driving cars', 'Google', 'few people', 'th
e company', 'him', 'I', 'you', 'very senior CEOs', 'major American car companies',
'my hand', 'I', 'Thrun', 'an interview', 'Recode']
Verbs: ['start', 'work', 'drive', 'take', 'tell', 'shake', 'turn', 'be', 'talk', '
say']
Sebastian Thrun PERSON
2007 DATE
American NORP
Thrun PERSON
Recode PERSON
earlier this week DATE
```

Features

- ✓ Support for **69+ languages**
- ✓ **58 trained pipelines** for 18 languages
- ✓ Multi-task learning with pretrained **transformers** like BERT
- ✓ Pretrained **word vectors**
- ✓ State-of-the-art speed
- ✓ Production-ready **training system**
- ✓ Linguistically-motivated **tokenization**
- ✓ Components for **named entity** recognition, part-of-speech tagging, dependency parsing, sentence segmentation, **text classification**, lemmatization, morphological analysis, entity linking and more
- ✓ Easily extensible with **custom components** and attributes
- ✓ Support for custom models in **PyTorch**, **TensorFlow** and other frameworks
- ✓ Built in **visualizers** for syntax and NER
- ✓ Easy **model packaging**, deployment and workflow management
- ✓ Robust, rigorously evaluated accuracy

Features

In the documentation, you'll come across mentions of spaCy's features and capabilities. Some of them refer to linguistic concepts, while others are related to more general machine learning functionality.

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) Tagging	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

NLTK Corpora

NLTK has built-in support for dozens of corpora and trained models, as listed below. To use these within NLTK we recommend that you consult the README file included with each corpus for further information.

NLTK 3.5 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

[Unicode Version 7.0.0 character properties in Perl](#) [[download](#) | [source](#)]
 id: unicode; size: 100266; author: ; copyright: ; license: ;

[aligner \(Sultan et al. 2015\) subset of the Paraphrase Database](#). [[download](#) | [source](#)]
 id: aligner; size: 711; author: ; copyright: ; license: Creative Commons Attribution 3.0 Unported (CC-BY);

[\[download | source \]](#)
 id: 3; author: Jan Strunk; copyright: ; license: ;

[Corpus de Sufixos da Língua Portuguesa](#) [[download](#) | [source](#)]
 id: portuguese; size: 3; author: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;

[es](#) [[download](#) | [source](#)]
 id: es; size: 100510; author: ; copyright: ; license: ;

[\[download | source \]](#)
 id: ; size: 6785405; author: ; copyright: ; license: ;

[\[download | source \]](#)
 id: ; size: 13404747; author: ; copyright: ; license: ;

[\[download | source \]](#)
 id: ; size: 10961490; author: ; copyright: ; license: ;

[\[download | source \]](#)
 id: del; size: 24516205; author: ; copyright: ; license: ;

[\[download | source \]](#)
 id: ; size: 49396025; author: ; copyright: ; license: ;

11. [Evaluation data from WMT15](#) [[download](#) | [source](#)]
 id: wmt15_eval; size: 383096; author: ; copyright: ; license: ;
12. [Grammars for Spanish](#) [[download](#) | [source](#)]
 id: spanish_grammars; size: 4047; author: Kepa Sarasola; copyright: ; license: ;
13. [Sample Grammars](#) [[download](#) | [source](#)]
 id: sample_grammars; size: 20293; author: ; copyright: ; license: ;
14. [Large context-free and feature-based grammars for parser comparison](#) [[download](#) | [source](#)]
 id: large_grammars; size: 283747; author: ; copyright: ; license: See the individual grammar files;
15. [Grammars from NLTK Book](#) [[download](#) | [source](#)]
 id: book_grammars; size: 9103; author: Ewan Klein; copyright: ; license: ;
16. [Grammars for Basque](#) [[download](#) | [source](#)]
 id: basque_grammars; size: 4704; author: Kepa Sarasola; copyright: ; license: ;

Some simple things you can do with NLTK

Tokenize and tag some text:

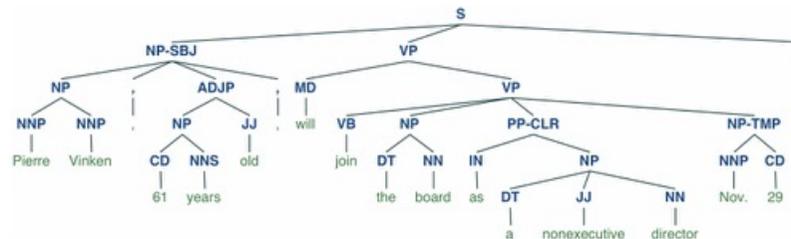
```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', 'o'clock', 'on', 'Thursday', 'morning',
'Arthur', 'did', 'n't', 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ('o'clock', 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

Identify named entities:

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [(('At', 'IN'), ('eight', 'CD'), ('o'clock', 'JJ'),
('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN')),
Tree('PERSON', [(('Arthur', 'NNP')]),
('did', 'VBD'), ('n't', 'RB'), ('feel', 'VB'),
('very', 'RB'), ('good', 'JJ'), (',', ','))])
```

Display a parse tree:

```
>>> from nltk.corpus import treebank
>>> t = treebank.parsed_sents('wsj_0001.mrg')[0]
>>> t.draw()
```



```
In [2]: from nltk.twitter import Twitter
tw = Twitter()
tw.tweets(keywords='love, hate', limit=10) #sample from the public stream
```

Sana magkakaisa na ang mga Kapamilya at Kapuso. Spread love, not hate
 #ShowtimeKapamiIyaDay #ALDubEBforLOVE
 @Real_Liam_Payne Please follow me , you mean the world to me and words can't describe how much i love you x3186
 Love my ugly wife
 RT @ansaberano: We Found Love
 #PushAwardsLizQuen
 RT @yungunmei: people want to fall in love but don't understand the concept
 I don't care, I love It #EMABiggestFans1D
 RT @bryan white: I'm not in the Philippines Yet but we are making a very BIG announcement in 2 days! Get ready! Love
 you! #GGMY #ALDubEBfor...
 I whole heartedly HATE @lakiamichelle like really HATE her 😡 who wants to be her friend because I DONT
 RT @lahrose23: I love yu to https://t.co/dfsRwSp1IC
 RT @alone_in_woods: ahoj, já jsem tvůj pes a tohle je náš love song /// Zrní - Já jsem tvůj pes https://t.co/7L0XPHeA
 2d via @YouTube
 Written 10 Tweets

Sentiment Analysis

```
>>> from nltk.classify import NaiveBayesClassifier
>>> from nltk.corpus import subjectivity
>>> from nltk.sentiment import SentimentAnalyzer
>>> from nltk.sentiment.util import *

>>> n_instances = 100
>>> subj_docs = [(sent, 'subj') for sent in subjectivity.sents(categories='subj')[n_instances]]
>>> obj_docs = [(sent, 'obj') for sent in subjectivity.sents(categories='obj')[n_instances]]
>>> len(subj_docs), len(obj_docs)
(100, 100)
```

Each document is represented by a tuple (sentence, label). The sentence is tokenized, so it is represented by a list of strings:

```
>>> subj_docs[0]
(['smart', 'and', 'alert', ',', 'thirteen', 'conversations', 'about', 'one',
 'thing', 'is', 'a', 'small', 'gem', '.'], 'subj')
```

We separately split subjective and objective instances to keep a balanced uniform class distribution in both train and test sets.

```
>>> train_subj_docs = subj_docs[:80]
>>> test_subj_docs = subj_docs[80:100]
>>> train_obj_docs = obj_docs[:80]
>>> test_obj_docs = obj_docs[80:100]
>>> training_docs = train_subj_docs+train_obj_docs
>>> testing_docs = test_subj_docs+test_obj_docs

>>> sentim_analyzer = SentimentAnalyzer()
>>> all_words_neg = sentim_analyzer.all_words([mark_negation(doc) for doc in training_docs])
```

We use simple unigram word features, handling negation:

```
>>> unigram_feats = sentim_analyzer.unigram_word_feats(all_words_neg, min_freq=4)
>>> len(unigram_feats)
83
>>> sentim_analyzer.add_feat_extractor(extract_unigram_feats, unigrams=unigram_feats)
```

We apply features to obtain a feature-value representation of our datasets:

```
>>> training_set = sentim_analyzer.apply_features(training_docs)
>>> test_set = sentim_analyzer.apply_features(testing_docs)
```

We can now train our classifier on the training set, and subsequently output the evaluation results:

```
>>> trainer = NaiveBayesClassifier.train
>>> classifier = sentim_analyzer.train(trainer, training_set)
Training classifier
>>> for key,value in sorted(sentim_analyzer.evaluate(test_set).items()):
...     print('{0}: {1}'.format(key, value))
Evaluating NaiveBayesClassifier results...
Accuracy: 0.8
F-measure [obj]: 0.8
F-measure [subj]: 0.8
Precision [obj]: 0.8
Precision [subj]: 0.8
Recall [obj]: 0.8
Recall [subj]: 0.8
```

<http://www.nltk.org>

Des bibliothèques Python (ou Java, C++, Swift)



Welcome to Apache OpenNLP

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.



scikit-learn

Machine Learning in Python

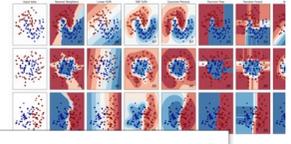
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...

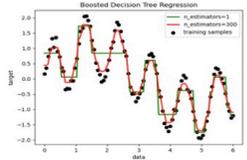


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...



CoreNLP



TensorFlow

Google is committed to

Une plate-forme Open Source de bout en bout dédiée au machine learning



Keras

Simple. Flexible. Powerful.

[Get started](#) [Guides](#) [API docs](#)

```
from tensorflow import keras
from tensorflow.keras import layers

# Instantiate a trained vision model
vision_model = keras.applications.ResNet50()

# This is our video-encoding pipeline using the trained vision_model
video_input = keras.Input(shape=(100, None, 3))
encoded_frame_sequence = layers.TimeDistributed(vision_model)(video_input)
encoded_video = layers.LSTM(256)(encoded_frame_sequence)

# This is our question-encoding pipeline using the trained question_model
question_input = keras.Input(shape=(100, ), dtype='int32')
embedded_question = layers.Embedding(10000, 256)(question_input)
encoded_question = layers.LSTM(256)(embedded_question)

# We then concatenate the encoded video and question
merged = keras.layers.concatenate([encoded_video, encoded_question])
output = keras.layers.Dense(1000, activation='softmax')(merged)
video_qn_model = keras.Model([video_input, question_input], output)
```

Deep learning for humans.

Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides.



PyTorch

Get Started Ecosystem Mobile Blog

FROM RESEARCH TO PRODUCTION

An open source machine learning framework that accelerates the path from research prototyping to production deployment.



Allen Institute for AI



AllenNLP

A natural language processing platform for building state-of-the-art models.



Des plateformes logicielles orientée « fouille de textes »

04.06.19

De nombreux acteurs industriels



Transformation digitale des entreprises - Xerox

À propos | Services | Produits | Fournitures | Support Client | Où acheter

Partager
Twitter
Lien
Courrier électronique
Imprimer
Plus

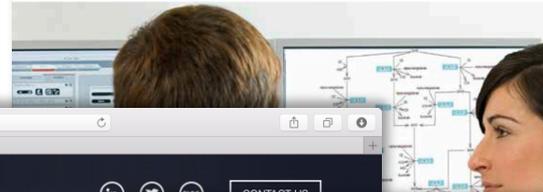
Analyse documentaire : l'arme secrète pour être en première ligne de la transformation numérique

Pour la plupart des entreprises, les ambitions numériques restent lettres mortes, comme le montre une enquête Xerox réalisée auprès de responsables IT. Le désir d'abandonner le papier pour



Elsevier R&D Solutions
FOR PHARMA & LIFE SCIENCES

Pathway Studio® Fact Sheet

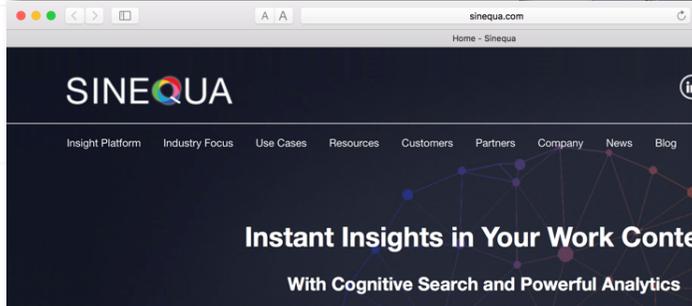


Linguamatics
an IQVIA company

Technology Products Solu



AI Siblings: NLP and Machine Learning for Better Drug Discovery



SINEQUA

Insight Platform | Industry Focus | Use Cases | Resources | Customers | Partners | Company | News | Blog

Instant Insights in Your Work Conte

With Cognitive Search and Powerful Analytics



Synapse

BESOINS SOLUTIONS PR

MONDECA

MAKING SENSE OF CONTENT

COMPRENDRE ET INTERPRETER

Comprendre le texte, l'image, la vidéo. Identifier ce qui est significatif.
Catégoriser. Evaluer. Trouver des relations. Désambigüiser. Enrichir.
Annoter. Gérer la connaissance pour nourrir les solutions d'intelligence artificielle: Classification automatique, bots, gestion d'alertes.

SYNAPSE DÉVELOPPEMENT

Les experts de l'Intelligence Artificielle appliquée au texte

- RÉINVENTER ma relation client
- VALORISER mes contenus
- OPTIMISER mes écrits

Nous utilisons les cookies pour vous offrir la meilleure expérience sur ce site. En continuant votre visite vous acceptez cette utilisation. [En savoir plus](#)



syllabs

Accueil Immobilier Nos offres d'emploi Contact EN ES

Entrez dans le futur de la création de contenu

Syllabs propose des solutions automatisées de création de textes et d'optimisation de contenus.
Notre approche unique qui conjugue expertise humaine et intelligence artificielle permet de répondre aux besoins d'information de tous vos publics, de développer votre trafic et d'optimiser votre stratégie SEO.



« Avec Syllabs, ma stratégie de contenu prend son envol »

Devon Think

The DEVONtechnology consists of a small and efficient kernel and several abstraction layers. Its core functions include:

- Data handling
- Semantic and associative data processing
- Signal processing
- Fast statistical analysis
- Fuzzy logic

Some of the most likely applications for the DEVONtechnology are structured and unstructured databases:

- Knowledge bases
- Expert systems
- Search engines
- Table-of-content generators
- Instant data-mining
- Intelligent agents
- Thesauri and context-sensitive help

In particular, applications depending on processing human language benefit from the flexible technology foundation.

The screenshot displays the Devon Think application interface. The main window shows a document titled "Operative report 11/1/2007". The document content includes a table of medical records and a detailed pathology report. The interface is annotated with several red arrows and text boxes:

- Powerful Search:** Limit to Current Selection, Expand to Encompass All, Open Databases
- Other Documents Closely Related According to Text Content:** Points to a list of related documents in the right-hand pane.
- A Complete Medical Chronological:** Points to the "Chronology" folder in the left-hand sidebar.
- All Files in Client Folder Indexed Can Be Tagged w/ OpenMeta:** Points to the "Tags" section in the sidebar.
- Color labels visually identify surgical, lab and radiology:** Points to color-coded text in the document.
- Hyperlink to the specific page in a multi page PDF:** Points to a blue hyperlink in the document.
- Word Index for Current Selection:** Points to a search results pane on the right.

<https://www.devontechnologies.com/apps/devonthink>



Enhance your feedback insights with text analytics

Keatext is an AI-driven text analytics platform that instantly processes your unstructured feedback, enabling you to dive into the data and find key insights right away.

Take your feedback analysis to the next level with automatic comment grouping, opinion and sentiment analysis, advanced data visualization, and powerful trend detection.



Multichannel analysis

Upload and get a global view on your feedback data from reviews, emails, surveys, help desk tickets, and call centre logs.

Opinion and sentiment analysis

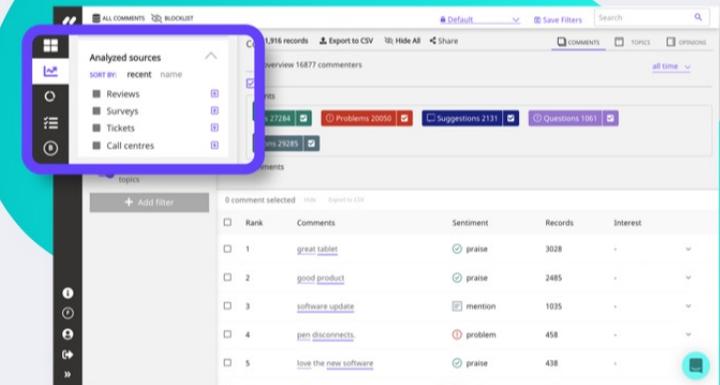
Instantly access the Praises, Problems, Suggestions, Questions, and Mentions in your feedback with accurate data categorization.

Multilingual processing

Analyze feedback with native proficiency in English and French, and top-of-the-line AI that translates 50+ other languages to English.

Advanced data filtering

Control and segment your data view with custom filters based on your metadata categories and relevant keywords.



rapidminer WHY RAPIDMINER INDUSTRIES PRODUCTS LEARN RESOURCES PARTNERS COMPANY

Depth for Data Scientists, Simplified for Everyone Else

Depth

- 1,500+ native algorithms, data prep & data science functions
- Support for any 3rd party ML libraries
- Notebooks & integration with custom Python & R
- Advanced analytics & powerful platform services

Augmented data science - guided & automated:

- Data cleansing & transformation
- Algorithm selection & validation
- Visual, intuitive model operations

Simplified

- Solution accelerators (pre-canned use case templates)
- Comprehensive tutorials
- Abstract complexity
- Self-paced online certification by persona
- Full automation where desired

opentext™ Solutions & Produits - Secteurs - Services - Support - Ressources - Contact France - Se connecter - Search

Accueil / Solutions et Produits / Découverte

Solutions de Discovery

Avec les logiciels et services eDiscovery et ses fonctions de machine learning, retrouvez plus rapidement les données clés, des données juridico-légales en passant par les données non structurées issues des analyses décisionnelles.

[Demande de démo](#) [Contactez-nous](#)

bin laden	official	intelligence	soldiers	palenets	arm
arnold	al gade				
11 016	polo	important	development	economic	middle
	middle east	investment	remark	agricultural	technology
	reducing	financial	countries	innovation	market
	global	growth			economic growth
11 023	rights	human rights	new york	human	freedom
	york	democracy	senat	activist	religion
	perit	courage	people	internet freedom	human rights activists
	new york times	honor			writing
11 021	software	software state gov	jackie software state	software jackie software	software
	christopher cam to s...	original message from...	message from software	jackie	parent
	gov	software jackie software	software state gov	state	software jackie
	state gov software state	software state gov su...			software state gov
11 018	minister	president	prime minister		
	paule	prime	other nations		

Group together similar documents by contextual theme to find relevant content

Hypergraph

Content - Sender/Recipient (To, CC, BCC) | untangled | dense | advanced

Legend: Start, Expanded, Start and Expanded, Selected

Nodes: h, mills, coleman, nora.toiv, unsubscribe, mills, cheryl d, jacob, colemancl, claire I, claire I subject: mini for ..., lon, sept 11, monica r, numa abedin, vulmore

Buttons: Reset, Narrow Search

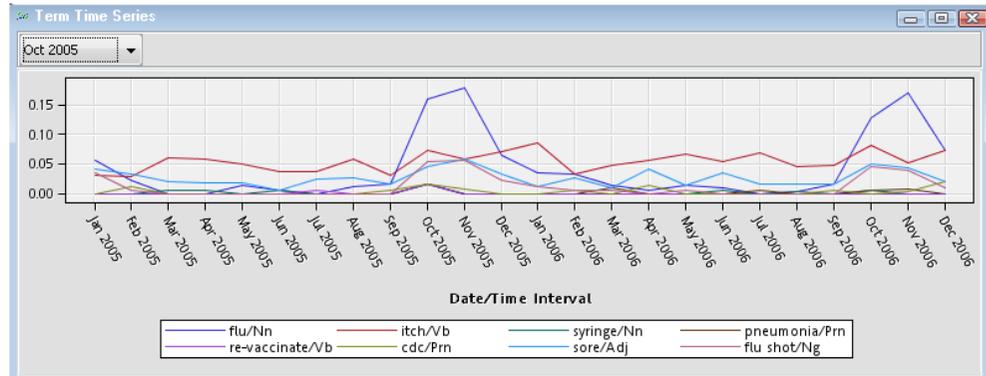
Document Date

Create maps of email and chat communications to quickly show who talked to whom



SAS® Text Miner

Mettez en évidence les informations dissimulées dans les données non structurées



Text Miner - Interactive

File Edit Tools View Window

Documents

SUMMARY

	NUMBER OF OCCURRENCES	VEHICLE/MAKE MAKE
REAR END COLLISION CAUSED SEAT BACKING TO COLLAPSE. *AK	0.0	TOYOTA TRUCK
WHILE WASHING TRUCK ACCIDENTLY RAISED HAND-OVER WINDSHIELD WIPER. REPLACEMENT (PYLON MANS) CUTTING HAND	1.0	TRUCK
WHILE DRIVING BRAKES LOCKED UP AND VEHICLE SLID, RESULTING IN A COLLISION. TOOK VEHICLE TO THE DEALER. ALSO, EXPERIENCED SOME VERY ERRATIC VIBRATION WHILE IN TRAFFIC. PLEASE PROVIDE ANY FURTHER INFORMATION. *AK	0.0	TOYOTA TRUCK
WHILE TRAVELING 35-35 MPH CONSUMER WAS BLIND BY ANOTHER VEHICLE'S FOG LIGHT (AND VEHICLE RAN INTO A TELEPHONE HEAD-ON), AIR BAGS DID NOT DEPLOY AT ANY TIME. MANUFACTURER HAS BEEN NOTIFIED. PLEASE PROVIDE FURTHER INFORMATION. *AK	0.0	TOYOTA TRUCK
WHILE PARKING VEHICLE IN PARK VEHICLE JUMPED OUT OF GEAR AND ACCELERATED FORWARD, RESULTING INTO AN ACCIDENT. THIS OCCURRED SEVERAL TIMES. *AK	0.0	SUBARU TRUCK
FIRST REAR SEAL, REAR SEAL AND BEARING, BEARING AND AXLE, NEW AXLE PUT IN THEN FLEW OFF TRUCK AND CRASHED, 3 WEEKS LATER REAR SEALS AXLE AGAIN HELP	4.0	TOYOTA TRUCK
VEHICLE WAS INVOLVED IN TWO ACCIDENTS IN WHICH THERE WAS NO DEPLOYMENT OF DRIVERS OR PASSENGERS SIDE AIRBAGS. THE FINAL ACCIDENT WAS A DIRECT FRONTAL IMPACT AT 40 MPH IN WHICH DRIVER WAS HURT. CAUSE OF PROBLEM UNKNOWN. *AK	2.0	TOYOTA TRUCK
UPON ACCELERATION AND MAKING A LEFT HAND TURN, STEERING WHEEL STUCK, CAUSING POOR STEERING CONTROL. (CONSUMER HAS CONTACTED DEALER, DEALER HAS YET TO DETERMINE THE CAUSE. PLEASE PROVIDE ANY FURTHER DETAILS. *AK	0.0	TOYOTA TRUCK
WHILE TRAVELING 35 MPH TRUCK SIMPLY GOT CAUGHT BETWEEN PAVEMENT GRABE ON SIDE OF ROAD. TRUCK THEN ROLLED TO RIGHT, AND OWNER TRIED TO MAINTAIN STABILITY. TRUCK WENT OFF THE ROAD, ROLLING-OVER SEVERAL TIMES. SEAT BELT BECAME LOOSE, AND NO AIR BAG DEPLOYED. THIS IS WAS A RENTAL VEHICLE. PLEASE DESCRIBE DETAILS. *AK	0.0	TOYOTA TRUCK
CONSUMER WAS INVOLVED IN A 35 MPH FRONTAL COLLISION IN WHICH THE DRIVERS/PASSENGERS AIR BAGS DID NOT DEPLOY. PLEASE GIVE ANY FURTHER DETAILS. *AK	0.0	PLYMOUTH TRUCK
WHILE STOPPING AT A TRAFFIC LIGHT VEHICLE WAS REAR ENDED AT 30 MPH, FORCED INTO ANOTHER VEHICLE HEAD-ON. UPON IMPACT, DRIVERS AND PASSENGERS AIR BAGS DID NOT DEPLOY. *AK. THE ACCIDENT CAUSED DAMAGE TO THE FRONT END OF THE VEHICLE AND INJURIES TO THE DRIVER. *YN	1.0	TOYOTA TRUCK

Clusters

#	Descriptive Terms	Freq	Percentage	PMS Std.
1	+ deploy, summary, above, re, + do, + hit, deployment, ar	273	0.0303642497941	0.002466558
2	hit, + number, hit number, + size, + tire, rear, tire stone, + control	309	0.0546554043967	0.007529618
3	acceleration, sudden, + accident, sudden acceleration, causing accident, + cause, reverse, abnorm	111	0.0155995502748	0.049955817
4	+ hit, off, + brake, + wheel, + side, + brake, rear, + cone	483	0.0606646306789	0.00037296
5	+ tire, hit, causing accident, + result, anti-lock, +a, fire, + accident	400	0.0621136480384	0.001822413
6	out of, down, + part, into, + jump, park, + gear, + rot	605	0.08501367337474	0.007910590
7	+ air, + stop, + brake, up, fire, with, when, + drive	2433	0.3483218426282	0.10741591
8	+ brake, + belt, + passenger, + injury, + seat, during, + impact, rear	503	0.0706577952726	0.009990240
9	+ apply, foot, + result, + brake, + stop, when, + failure, + hit	716	0.13061832400163	0.048320615
10	+ hit, + impact, re, + deploy, + collision, + side, + do, + re	1173	0.1448237918392	0.04722741

Terms

TERM	Freq	# Documents	Keep	VEIGHT	Role	Attribute
in	4128	3019	<input type="checkbox"/>	0.1112401	Prep	Alpha
not	3027	2385	<input type="checkbox"/>	0.1429807	Part	Alpha
brake	3056	2091	<input type="checkbox"/>	0.15329183	Noun	Alpha
on	2909	1806	<input type="checkbox"/>	0.16603	Prep	Alpha
accident	2005	1817	<input type="checkbox"/>	0.158637	Noun	Alpha
cause	1887	1729	<input type="checkbox"/>	0.1622358	Verb	Alpha
causing	1225	1196	<input type="checkbox"/>			
cause	51	46	<input type="checkbox"/>			
sound	509	500	<input type="checkbox"/>			
causers	32	21	<input type="checkbox"/>			
do	1900	1826	<input type="checkbox"/>	0.173788	Aux	Alpha
driver	1839	1480	<input type="checkbox"/>	0.1340025	Noun	Alpha
display	1403	1389	<input type="checkbox"/>	0.1574912	Verb	Alpha
high	1238	1192	<input type="checkbox"/>	0.2031426	Prop	Alpha
hit	1326	1174	<input type="checkbox"/>	0.2080041	Verb	Alpha
into	1321	1149	<input type="checkbox"/>	0.2118734	Prep	Alpha
when	1265	1146	<input type="checkbox"/>	0.2106032	Conj	Alpha
consumer	1507	1071	<input type="checkbox"/>	0.2274474	Noun	Alpha
air	1170	1046	<input type="checkbox"/>	0.2217276	Noun	Alpha
bag	1158	1027	<input type="checkbox"/>	0.2225795	Noun	Alpha

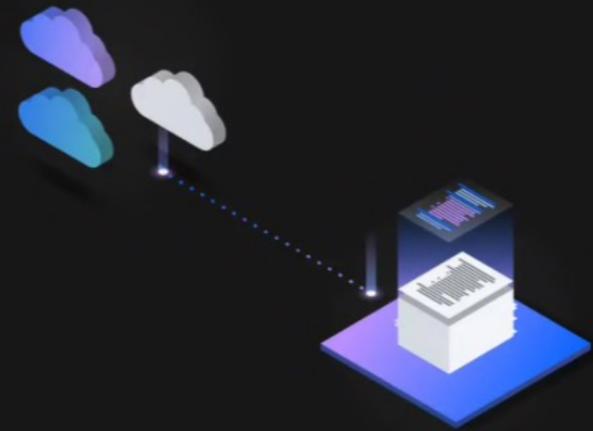
Concept Linking

Watson Natural Language Understanding

The natural language processing (NLP) service for advanced text analytics

[Get started free](#)

[→ View demo](#)



Site feedback

What is Watson Natural Language Understanding?

Watson Natural Language Understanding is a cloud native product that uses deep learning to extract metadata from text such as entities, keywords, categories, sentiment, emotion, relations, and syntax.

<h3>Overview</h3>	<h4>Powerful Insight Extraction</h4> <p>Get underneath the topics mentioned in your data by using text analysis to extract keywords, concepts, categories and more.</p> <p>Learn more</p>	<h4>Extensive Language Support</h4> <p>Analyze your unstructured data in more than thirteen languages.</p> <p>Learn more</p>	<h4>High-Accuracy Extraction</h4> <p>Out-of-the-box machine learning models for text mining provide a high degree of accuracy across your content.</p> <p>Learn more</p>
<h3>Why Watson NLU?</h3>	<h4>Deploy Anywhere</h4> <p>Deploy Watson Natural Language Understanding behind your firewall or on any cloud.</p> <p>Learn more</p>	<h4>Domain Customization</h4> <p>Train Watson to understand the language of your business and extract customized insights with Watson Knowledge Studio.</p> <p>Learn more</p>	<h4>Data Control</h4> <p>Maintain ownership of your data with the assurance that your data is safe and secure. IBM will not collect or store your data.</p> <p>Learn more</p>



[Let's talk](#)

SAMPLE INDUSTRY DOMAINS

Legal Financial

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**,

■ Entities (Out of the document)

Extraction

Entities Key

Name
Directors and Corporate Governance Committee
full Board
IBM

SAMPLE INDUSTRY DOMAINS

Legal Financial

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**,

■ Neutral Entity

Extraction

Sentiment Emotion

Full Document

Entity Sentiment

Directors and Corporate Governance Committee

full Board

IBM

Keyword Sentiment

IBM Board Corporate Governance Guidelines

independent directors

part of the assessment of director independence

Corporate Governance Committee

Directors

full Board

SAMPLE INDUSTRY DOMAINS | TRY YOUR OWN

Legal Financial Media | Input Text URL

Under the **IBM Board Corporate Governance Guidelines**, the **Directors and Corporate Governance Committee** and the **full Board** annually review the financial and other **relationships** between the **independent directors** and IBM as **part of the assessment of director independence**. The **Directors and Corporate Governance Committee** makes **recommendations** to the Board about the independence of non- **management directors**, and the Board determines whether these directors are independent. In **addition to this annual assessment of director independence**,

■ Sadness ■ Fear ■ Disgust ■ Anger ■ Joy

Extraction Classification Linguistics Custom

Sentiment Emotion Categories

Full Document

Emotion	Percentage
Sadness	1.73%
Joy	32.41%
Fear	3.74%
Disgust	4.37%
Anger	10.53%

Entity Emotion Scores

Directors and Corporate Governance Committee

Emotion	Percentage
Sadness	1.73%
Joy	32.41%
Fear	3.74%





Gargantext

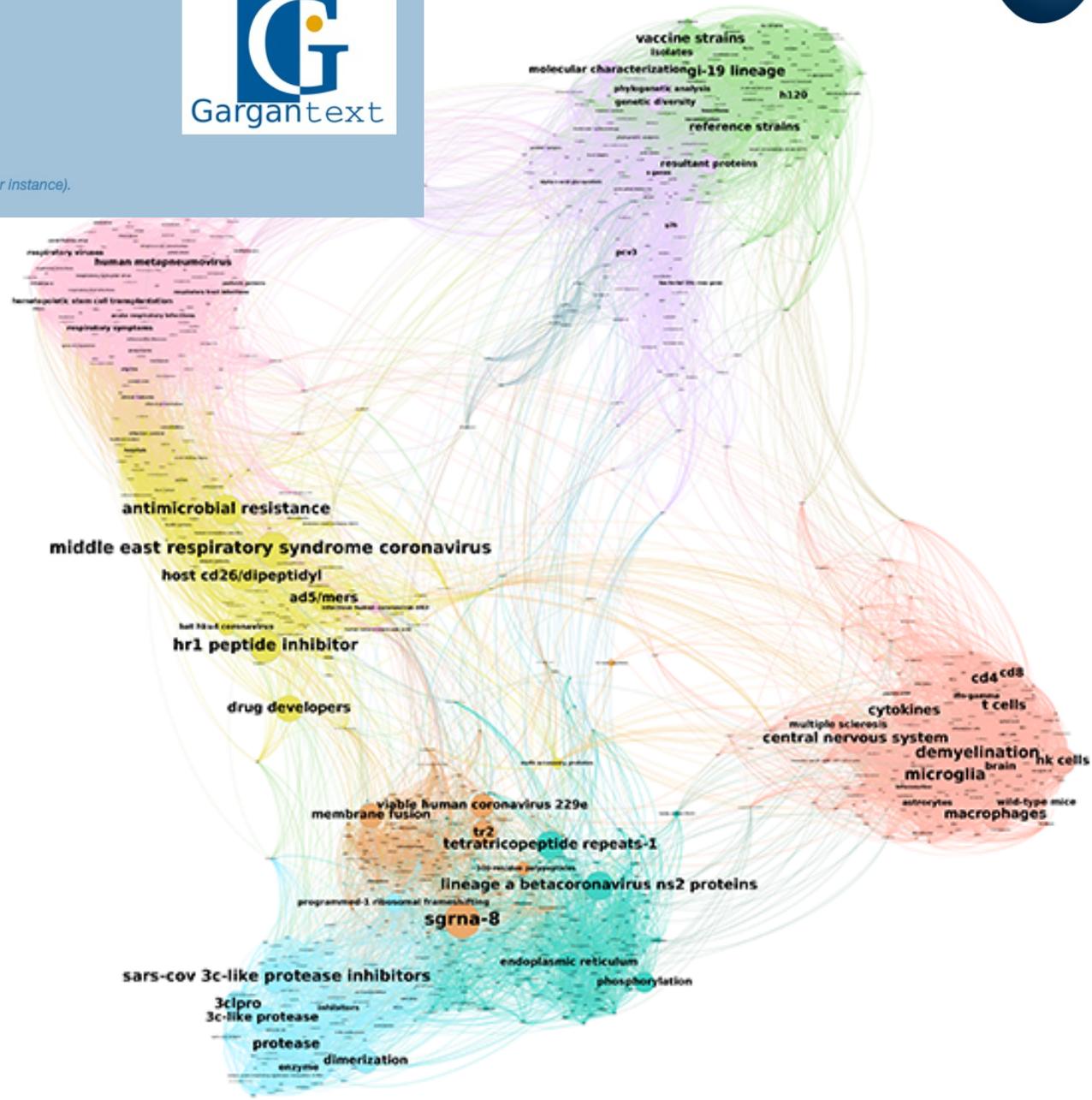
A web platform to explore text-mining

Log in

Sign Up

Documentation

Some features may not work without a javascript optimized browser (Chromium for instance).

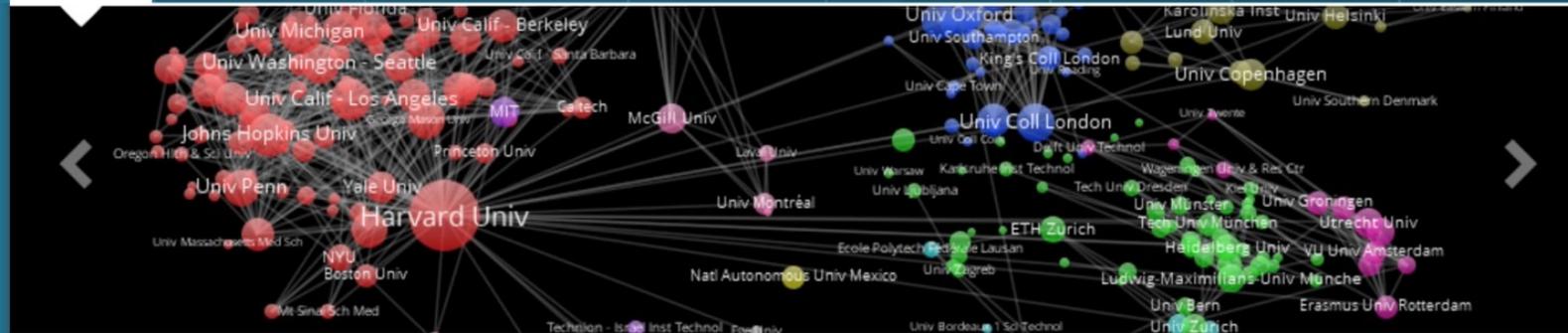


a collaborative web platform for the exploration of sets of unstructured documents. It combines tools from natural language processing, text-mining, complex networks analysis and interactive data visualization to pave the way toward new kinds of interactions with your digital corpora. In few minutes, you will be able to do knowledge maps, collaborative state-of-the-art, portfolio analysis

<https://gargantext.org>



<https://lejournal.cnrs.fr/articles/visualiser-la-recherche-sur-le-coronavirus-en-un-coup-doeil>



Welcome to VOSviewer

VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on co-citation, bibliographic coupling, or co-authorship relations. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature.

VOSviewer version 1.6.5

VOSviewer version 1.6.5 was released on September 28, 2016. Some of the improvements introduced in this version are listed below:

- **Overlay visualizations.** These popular visualizations have been made more prominently visible.
- **Maps based on bibliographic data.** Functionality for creating maps based on bibliographic data has been improved. Items can be filtered based on citation counts, and various types of overlay visualizations are supported.
- **Command line parameters.** Many

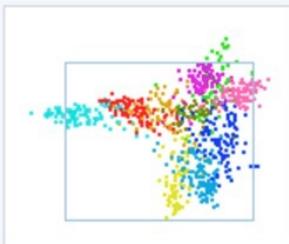


VOSviewer web start

Click the button below to launch VOSviewer directly from this web page. This requires a system with Java support.

[Launch VOSviewer](#)





Action Items Map

Filter:
schedul

16 items in 3 clusters:

Cluster 6 (14 items)

- batch scheduling
- feasible schedule
- flowshop scheduling problem
- job shop scheduling
- job shop scheduling problem
- line scheduling
- machine scheduling problem
- optimal schedule
- project scheduling
- project scheduling problem
- scheduling
- scheduling problem
- single machine scheduling
- single machine scheduling

Cluster 7 (1 item)

- production schedule

Cluster 9 (1 item)

- crew scheduling

< >

Group items by cluster

Network Visualization Density Visualization



Labels

Size:

Size variation:

Max. length:

Font:

Visualization

Item density

Cluster density

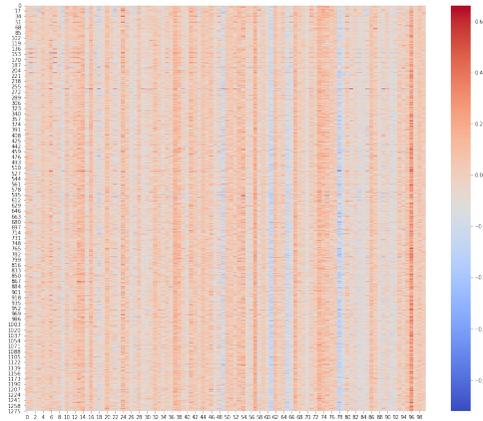
Kernel width:

Colors

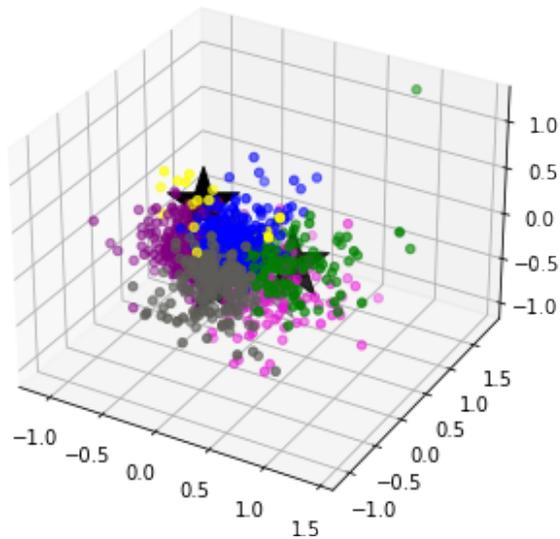
White background

Visualisation ?

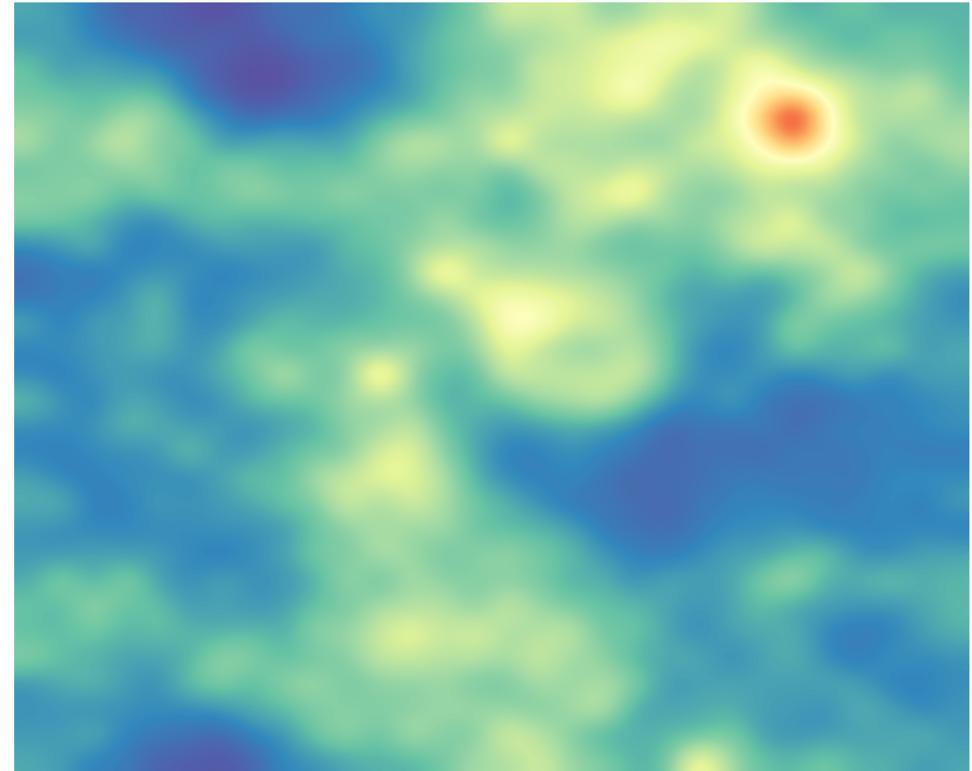
Représentation de documents par projection



Vectorisation dans des espaces continus (Doc2Vec) par plongements lexicaux



Analyse en Composantes Principales



Cartes auto-organisées (SOM)

Dörk, M., C. Williamson, and S. Carpendale. "Towards Visual Web Search: Interactive Query Formulation and Search Result Visualization." In *WSSP. Madrid, Spain, 2009.*



Search

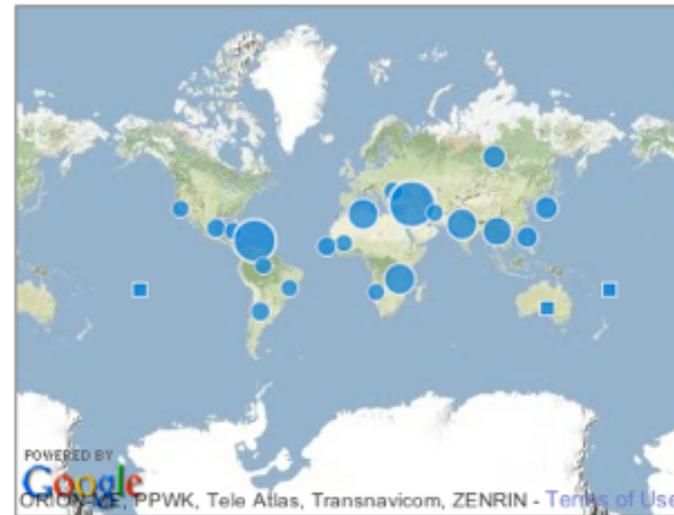
Collection

[About](#) [Help](#)

Time ✕



Location +



Tags

arts culture business children cyberactivism development diaspora disaster economics education elections energy entertainment environment ethnicity finance food freedom of speech gender governance health history humanitarian human rights humor ideas indigenous industry international relations internet telecoms labor language law LGBT literature media music photos politics protest racism religion runetechno software tools sport technology transparency and technology network travel video war conflict youth

Results

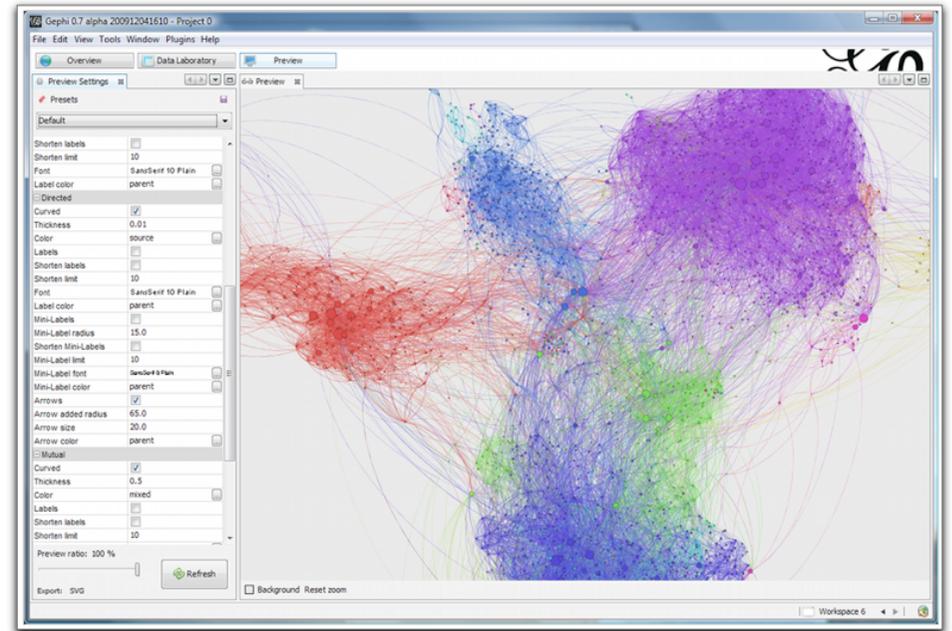
671 time title 1 2 3 4 5 6 7 8 9 10 11 12 13 14

- | | | | | | | | | | | | | | |
|---|--|---|---|---|---|--|---|----------------------------------|--|---|---|--|---|
| Angola: Similarities between Cabinda and East Timor | Zimbabwe: Is it a crime to be white in Zimbabwe? | Ghana: Ghana National Media Commission | Africa: A blog to promote African literature | South Africa Rugby: Playing Its Part in Transformation | Ecuador: Tungurahua Volcano Erupts and Spreads Ash | Nigeria: Exclusive interview with Nigerian soul artist | France: Questions and Controversy about ICC | Oil spill disaster in Singapore | Philippines: Destroying another landmark | Vietnam: Eating "Op la" in Saigon | Vietnam: Report on online censorship | Colombia: Tweeting the May 30 Presidential Elections | Guinea: Waiting for Presidential Elections |
| Mexico: 165 Mexicans Die Each Day Due to Smoking | Guatemala: Cleaning Up the Ash from Pacaya Volcano | France: A Legal Review of the Burqa Ban Bill | Brazil: Rapper assaulted by Police | India: Video Giving A Voice To Marginalised Communities | Caucasus: Eurovision Semi-final roundup | Bahrain: On German Freedom | Pakistan: Minorities At Peril | India: Second Class Rail Travel | Pakistan: Attacks in Lahore at Ahmadi Mosque | Algeria: "France: Film 'Of Gods and Men' Sparks | Bangladesh: FIFA World Cup: Memories From The | Zimbabwe: Outlawed newspaper coming back | Eritrea: Exiled editor reunites with family |
| Caucasus: Boobs, cleavage, and a rare unity in | Nepal: Financial Fraud | Russia: Blogger and Activist Arrested for Viral Video | Iran: Two Bloggers and Student Leader on Hunger | Guatemala: Pacaya Volcano Causes State of | Israel: The Freedom Flotilla - PR Stunt or Humanitarian | Hungary: Facebook vs. WW | Hungary: Bloggers' Photo and Video Reports on | Jamaica: Dudas, Security & Seaga | Lebanon: Elections Vs Football | Bahrain: Abu Dhabi's Gold ATMs | Jordan: A Day with a Grave Digger | MENA: 15m Facebook Users | Qatar: The Al Jazeera Initiative for Internet Freedom |

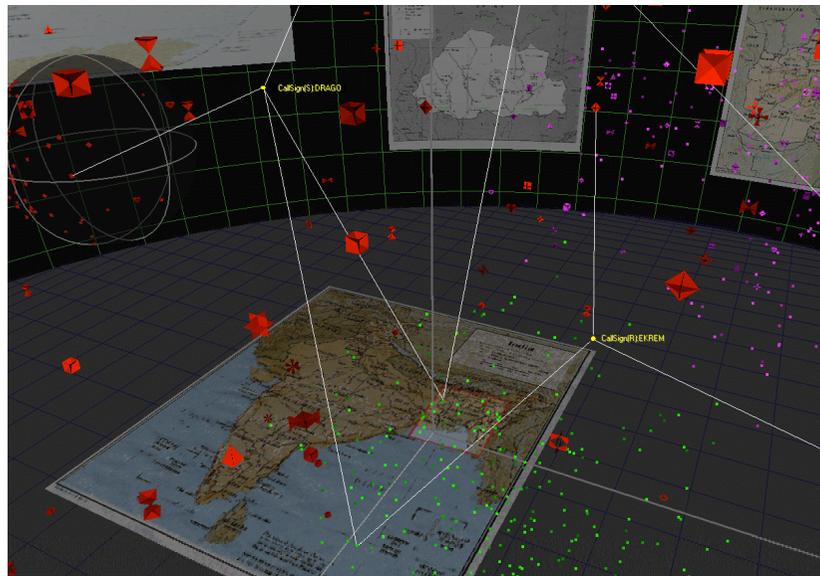




(L) LUCKYSORT



gephi CONSORTIUM



projector.tensorflow.org

word embedding explorer - Recherc... Embedding projector - visualization... wevi Vector explorers Word embedding demo

Embedding Projector

DATA Points: 10000 | Dimension: 200

5 tensors found
Word2Vec 10K

Label by
word

Color by
No color map

Sphereize data

Load data Publish

Checkpoint: Demo datasets
Metadata: oss_data/word2vec_10000_200d_labels.tsv

T-SNE **PCA** CUSTOM

X Component #1 Y Component #2
Z Component #3

PCA is approximate.
Total variance described: 8.5%

Show All Data Isolate 5948 points Clear selection

Search by word

<https://projector.tensorflow.org>

BOOKMARKS (0)

projector.tensorflow.org

word embedding explorer - Recherc... Embedding projector - visualization... wevi Vector explorers Word embedding demo

Embedding Projector

DATA Points: 10000 | Dimension: 200 | Selected 101 points

5 tensors found
Word2Vec 10K

Label by
word

Color by
No color map

Sphereize data

Load data **Publish**

Checkpoint: Demo datasets
 Metadata: oss_data/word2vec_10000_200d_labels.tsv

T-SNE **PCA** CUSTOM

X Component #1 Y Component #2
 Z Component #3

PCA is approximate.
 Total variance described: 8.5%

free
 word free
 count 5684

Search ee by word
 neighbors 100
 distance COSINE EUCLIDEAN

Nearest points in the original space:

open	0.578
freely	0.580
available	0.592
freedom	0.616
software	0.638
source	0.647
online	0.678
tools	0.686
independent	0.687
allow	0.689
content	0.690
complete	0.695
new	0.696
good	0.696
allowing	0.699
support	0.702
bound	0.706
access	0.709
create	0.710
licensing	0.711
thus	0.711
strong	0.718
download	0.719
developers	0.720
fair	0.722

BOOKMARKS (0)

Des tâches pour la recherche en informatique

04.06.19

DES APPLICATIONS

- Enrichissement ou analyse d'un document :
 - Identification d'entités dans les textes (noms propres, dates, lieux...)
 - Classification et catégorisation automatiques
 - Construction de terminologie, extraction de mots-clés
 - Identification de citations et structuration de références bibliographiques
 - Structuration de documents « image » et reconnaissance de caractères
 - Analyse de sentiment
 - Résumé automatique
 - Attribution d'auteur
- Intégration de plusieurs niveaux d'analyse sur des collections)
 - Recherche de documents, d'images, de pages Web
 - Recommandation automatique de contenus, veille
 - Systèmes de questions-réponses
 - Cartographie et navigation guidées
 - Détection de nouveauté,, de tendances...

De nombreuses compétitions



The CLEF Initiative
Conference and Labs of the Evaluation Forum

<http://www.clef-initiative.eu/>

1. **ARQMath: Answer Retrieval for Questions on Math**
2. **BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering**
3. **CheckThat!: Automatic Identification and Verification of Claims**
4. **ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents**
5. **eHealth: Retrieving and Making Sense of Medical Content**
6. **eRisk: Early Risk Prediction on the Internet**
7. **HIPE: Identifying Historical People, Places and other Entities**
8. **ImageCLEF: Multimedia Retrieval in Medicine, Lifelogging, and Internet**
9. **LifeCLEF: Multimedia Retrieval in Nature**
10. **LiLAS: Living Labs for Academic Search**
11. **PAN: Stylometry and Digital Text Forensics**
12. **Touché: Argument Retrieval**

LifeCLEF - Biodiversity identification and prediction

ProtestNews - Extracting Protests from News

eHealth

CENTRE@CLEF

eRISK - Early Risk prediction on the Internet

PAN Lab on Digital Text Forensics and Stylometry

CheckThat! - Automatic Identification and Verification of Claims

PIR-CLEF - Evaluation of personalised IR

SemEval-2021

The 15th International Workshop on Semantic Evaluation

<https://semeval.github.io/SemEval2021/tasks.html>

Lexical semantics

- **Task 1: Lexical Complexity Prediction** ([email organizers] [mailing list])
Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, Marcos Zampieri
- **Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation** ([email organizers])
NOTE: new competition website!
Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli
- **Task 3: Span- and Dependency-based Multilingual and Cross-lingual Semantic Role Labeling**
- **Task 4: Reading Comprehension of Abstract Meaning** ([email organizers] [mailing list])
Boyuan Zheng, Xiaoyu Yang, Yu-Ping Ruan, Quan Liu, Zhen-Hua Ling, Si Wei, Xiaodan Zhu

Social factors & opinion

- **Task 5: Toxic Spans Detection** ([email organizers] [mailing list])
John Pavlopoulos, Ion Androutsopoulos, Jeffrey Sorensen, Léo Laugier
- **Task 6: Detection of Persuasive Techniques in Texts and Images** *Updated website* ([email organizers] [mailing list])
Giovanni Da San Martino, Hamed Firooz, Preslav Nakov, Fabrizio Silvestri
- **Task 7: HaHackathon: Detecting and Rating Humor and Offense** ([email organizers])
NOTE: new competition website!
J. A. Meaney, Steven Wilson, Walid Magdy, Luis Chiruzzo

Information in scientific & clinical text

- **Task 8: MeasEval: Counts and Measurements** ([email organizers] [mailing list])
Corey Harper, Jessica Cox, Ron Daniel, Paul Groth, Curt Kohler, Antony Scerri
- **Task 9: Statement Verification and Evidence Finding with Tables** ([email organizers] [mailing list])
Nancy Xin Ru Wang, Sara Rosenthal, Marina Danilevsky, Diwakar Mahajan
- **Task 10: Source-Free Domain Adaptation for Semantic Processing** ([email organizers] [mailing list])
Steven Bethard, Egoitz Laparra, Timothy Miller, Özlem Uzuner
- **Task 11: NLPContributionGraph** ([email organizers] [mailing list])
Jennifer D'Souza, Sören Auer, Ted Pedersen

De nombreuses compétitions

The screenshot shows the Kaggle interface for the 'Getting Started Prediction Competition' titled 'Natural Language Processing with Disaster Tweets'. The main text reads: 'Predict which Tweets are about real disasters and which ones are not'. It is organized by Kaggle, with 1,325 teams and is ongoing. The navigation menu includes Overview, Data, Code, Discussion, Leaderboard, Datasets, and Rules. The description section is partially visible, starting with 'Welcome to one of our "Getting Started" competit'.

<https://www.kaggle.com/c/nlp-getting-started>

The screenshot shows the CodaLab Competitions website. The URL is <https://competitions.codalab.org/competitions/>. The page features a search bar and a list of competitions:

- Evaluating grammatical error corrections**: Organized by cnapoles. This "competition" contains different evaluation metrics commonly used for GEC and allows users to score their systems with these metrics ...
- ADoBo — Automatic Detection of Borrowings**: Organized by lea. Detecting emerging borrowings from English into Spanish (words like 'smartphone' or 'fake news') that appear in the Spanish press
- Interspeech Shared Task: Automatic Speech Recognition for Non-Native Children's Speech**: Organized by cleong. A joint shared task between FBK, ETS and Cambridge
- ICDAR 2021 Competition on Scene Video Text Spotting**: Organized by Embers. To support this competition, we extend the Larger-Scale Video Text Dataset released in YORO [1], and release a dataset containing ...
- EmoEvalEs@IberLEF 2021**: Organized by amontejo. Workshop en Emotion detection and Evaluation for Spanish



Thirteenth Text Analysis Conference (TAC 2020)

Evaluation: August 2020 - January 2021
Workshop: February 22-23, 2021

- Epidemic Question Answering (EPIC-QA)**
 The goal of the EPIC-QA track is to evaluate systems on their ability to answer questions about COVID-19, related coronaviruses, and the recommended response to the pandemic. The goal is to return expert-level answers as expected by the scientific community.
Track coordinators: Dina Demner-Fushman (ddemner@mail.nih.gov)
Home page: https://bionlp.nlm.nih.gov/epic_qa/
Group / mailing list: epic-qa@list.nist.gov
- Recognizing Ultra Fine-grained Entities (RUFES)**
 The goal of the KBP RUFES track is to extract and corefer mentions of entities.
Track coordinator: Heng Ji (hengji@illinois.edu) and Avirup Sil (avirup@illinois.edu)
Home page: <https://tac.nist.gov/2020/KBP/RUFES/>
Group / mailing list: tac-kbp@list.nist.gov
- Streaming Multimedia Knowledge Base Population (SM-KBP)**
 The goal of the SM-KBP track is to develop and evaluate technologies for populating a knowledge base with explicit alternative interpretations of events, situations, and trends in multimedia.
Track coordinator: Hoa Dang (hoa.dang@nist.gov)
Home page: <https://tac.nist.gov/2020/KBP/SM-KBP/>
Group / mailing list: sm-kbp@list.nist.gov

<https://tac.nist.gov/2020/index.html>



De nombreuses compétitions

DEFT (DÉfi Fouille de Textes)

<https://deft.limsi.fr>

- **2005** (*Dourdan, France, TALN 2005*) : identification du locuteur d'un discours politique parmi deux protagonistes différents (Jacques Chirac vs. François Mitterrand).
- **2006** (*Fribourg, Suisse, SDN 2006*) : segmentation thématique de textes politiques.
- **2007** (*Grenoble, France, AFIA 2007*) : détection de l'opinion exprimée dans un texte de retranscription de débats parlementaires (projets de Loi relatifs à l'énergie).
- **2008** (*Avignon, France, TALN 2008*) : classification automatique de documents en genres (*journalistique vs. encyclopédiques*) et thèmes différents (*art, économie, littérature, politique internationale, politique nationale, problèmes de sociétés, sciences, sports, télévision*).
- **2009** (*Paris, France*) : fouille d'opinion (objectif/subjectif) en corpus multilingues (journaux et débats européens).
- **2010** (*Montréal, Canada, TALN 2010*) :
 - Variation diachronique (1800-1944) en corpus de presse française (*Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, La Croix, Le Figaro*), identification de la décennie de publication d'un extrait d'article ;
 - Variation diatopique en corpus de presse française (*L'Est Républicain, Le Monde*) et québécoise (*La Presse, Le Devoir*).
- **2011** (*Montpellier, France, TALN 2011*) :
 - Variations diachroniques (1800-1944) en corpus de presse française (*Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, La Croix, Le Figaro, La Presse, Le Temps*), identification de l'année de publication d'un extrait d'article ;
 - Appariements résumé/article scientifique de revue dans le domaine des Sciences Humaines et Sociales (Humanités).
- **2012** (*Grenoble, France, TALN 2012*) : identification automatique des mots-clés indexant le contenu d'articles scientifiques ayant paru en revues de Sciences Humaines et Sociales, avec l'aide de la terminologie des mots-clés (piste 1), sans terminologie (piste 2).
- **2013** (*Les Sables-d'Olonne, France, TALN 2013*) : identification du niveau de difficulté de réalisation d'une recette, identification du type de plat préparé, appariement d'une recette avec son titre, identification des ingrédients d'une recette.
- **2014** (*Marseille, France, TALN 2014*) : catégoriser le genre littéraire de courtes nouvelles, évaluer la qualité littéraire de ces nouvelles, déterminer si une œuvre fait consensus, déterminer la session scientifique dans laquelle un article de conférence a été présenté.
- **2015** (*Caen, France, TALN 2015*) : fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter.
- **2016** (*Paris, France, TALN 2016*) : indexation de documents scientifiques en français.
- **2017** (*Orléans, France, TALN 2017*) : fouille d'opinion dans des messages postés sur Twitter.
- **2018** (*Rennes, France, CORIA-TALN 2018*) : recherche d'information et analyse de sentiments dans des tweets sur les transports en Ile-de-France.
- **2019** (*Toulouse, France, PFIA-TALN-RECITAL 2019*) : recherche et extraction d'information dans des cas cliniques
- **2020** (*Nancy, France, conférence virtuelle JEP-TALN-RECITAL 2020*) : similarité sémantique et extraction d'information fine dans des cas cliniques

CONCLUSION : LE TDM...

Nécessite :

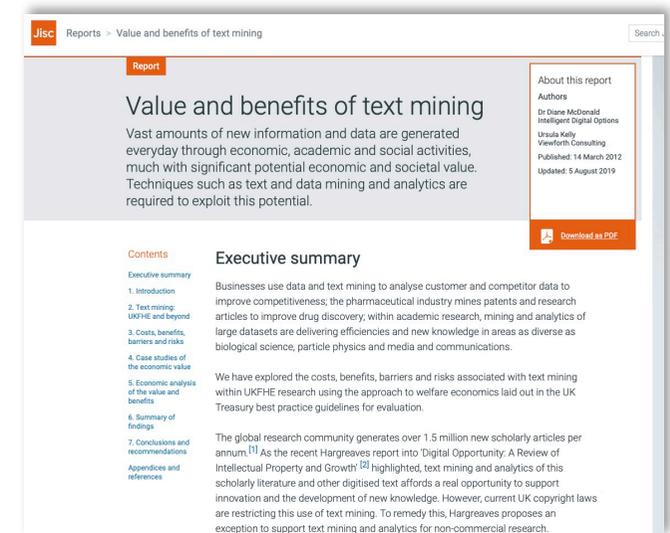
- un corpus cible, des ressources de spécialité
- d'intégrer différents composants logiciels
- un scénario et une référence pour évaluer la chaîne de traitements

Faisable si :

- les composants sont interopérables, les métadonnées compatibles
- l'intégration de différents composants logiciels est possible (ou s'il existe déjà une brique logicielle répondant au besoin)

Concerne et impacte :

- La recherche scientifique dans son ensemble
- La société au travers d'applications du quotidien



The screenshot shows the Jisc website page for the report 'Value and benefits of text mining'. The page includes a navigation bar with 'Jisc Reports > Value and benefits of text mining' and a search box. The main content area features the report title, a brief description, and a list of authors: Dr Diane McDonald (Intelligent Digital Systems) and Ursula Kelly (Viewforth Consulting). It also displays the publication date (14 March 2012) and the last update date (5 August 2019). A 'Download as PDF' button is visible. The 'Contents' section lists: Executive summary, 1. Introduction, 2. Text mining: UKFHE and beyond, 3. Costs, benefits, barriers and risks, 4. Case studies of the economic value, 5. Economic analysis of the value and benefits, 6. Summary of findings, 7. Conclusions and recommendations, and Appendices and references. The 'Executive summary' text states that businesses use data and text mining to analyse customer and competitor data to improve competitiveness, and that the pharmaceutical industry mines patents and research articles to improve drug discovery. It also mentions that large datasets are delivering efficiencies and new knowledge in areas as diverse as biological science, particle physics and media and communications. The summary further notes that the global research community generates over 1.5 million new scholarly articles per annum, and that text mining and analytics of this scholarly literature and other digitised text affords a real opportunity to support innovation and the development of new knowledge. However, current UK copyright laws are restricting this use of text mining. To remedy this, Hargreaves proposes an exception to support text mining and analytics for non-commercial research.

<https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>