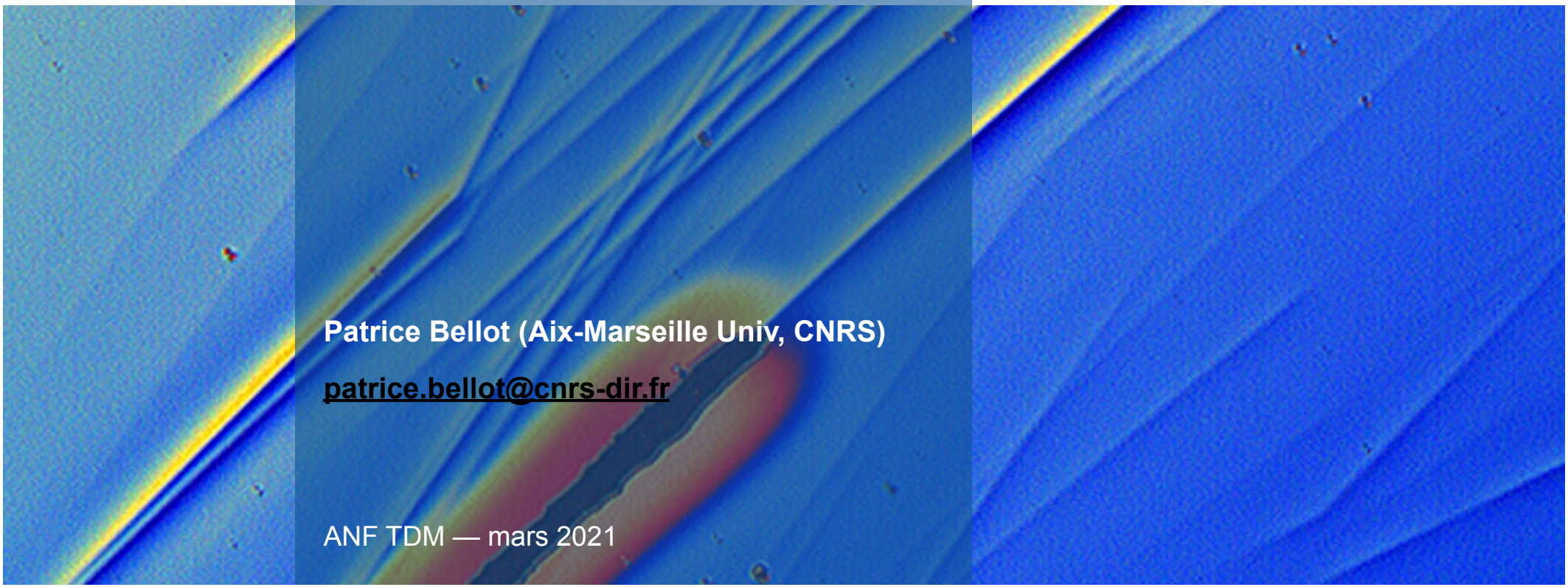
Introduction à la FOUILLE DE TEXTES et positionnement de l'offre logicielle



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

Des difficultés génériques.

Des standards nécessaires.

Des solutions à partager.

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


Briefing

Requested by the JURI committee

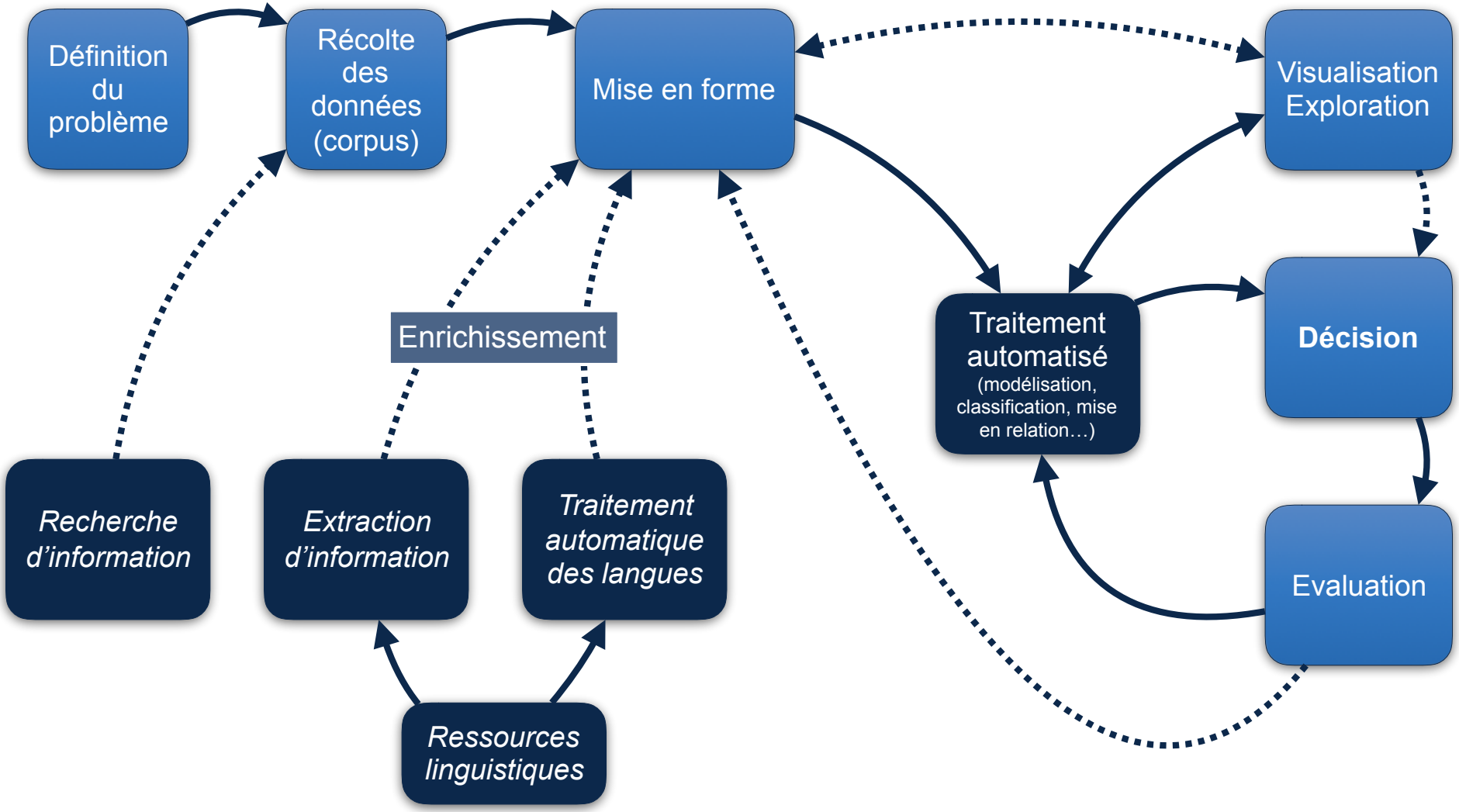
European Parliament

The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects

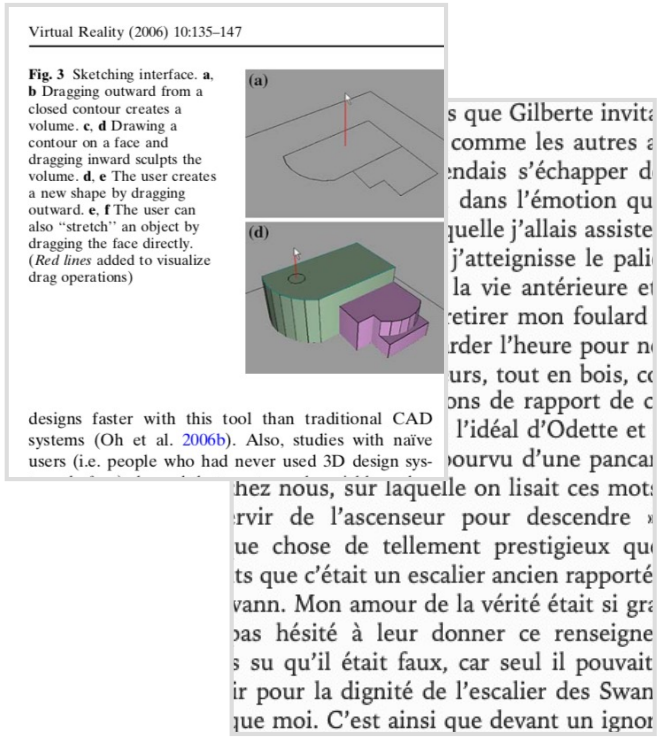
loi française Pour une République Numérique du 7 octobre 2016 et la directive européenne sur le droit d'auteur dans le marché unique numérique du 26 mars 2019

Des données, des méta-données et des formats

Un processus de fouille de textes



Des données



Virtual Reality (2006) 10:135–147

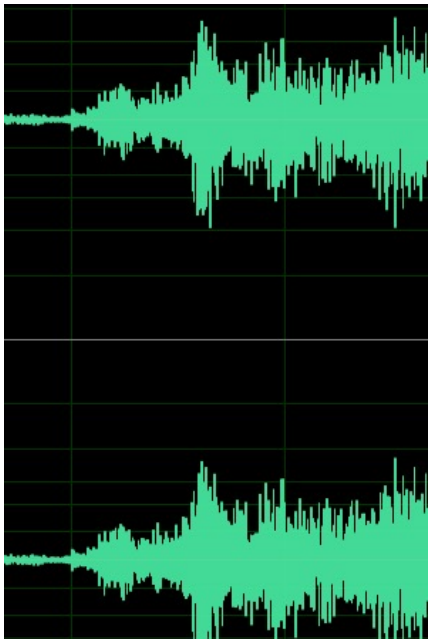
Fig. 3 Sketching interface. **a**, **b** Dragging outward from a closed contour creates a volume. **c**, **d** Drawing a contour on a face and dragging inward sculpts the volume. **d**, **e** The user creates a new shape by dragging outward. **e**, **f** The user can also "stretch" an object by dragging 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
ndais s'échapper d
dans l'émotion qu
uelle j'allais assiste
j'atteignisse le pali
la vie antérieure et
etirer mon foulard
rder l'heure pour n
urs, tout en bois, c
ons de rapport de c
l'idéal d'Odette et
ourvu d'une pancar
hez nous, sur laquelle on lisait ces mot
rvir de l'ascenseur pour descendre »
ue chose de tellement prestigieux qu
ts que c'était un escalier ancien rapporté
vann. Mon amour de la vérité était si gra
as 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
ue moi. C'est ainsi que devant un ignor

Texte

Image



Audio

Des données « numérisables »

s, certaines des amies que Gilberte invita
ıt obligées de partir comme les autres a
, dès l'escalier j'entendais s'échapper d
urmure de voix qui, dans l'émotion qu
ıonie imposante à laquelle j'allais assiste
ent, bien avant que j'atteignisse le palie
ttachaient 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, cc
dans certaines maisons de rapport de c
vait été si longtemps l'idéal d'Odette et
ôt se déprendre, et pourvu d'une pancar
hez nous, sur laquelle on lisait ces mots
ervir de l'ascenseur pour descendre »
ue chose de tellement prestigieux que
ıts que c'était un escalier ancien rapporté
vann. Mon amour de la vérité était si gra
as 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)

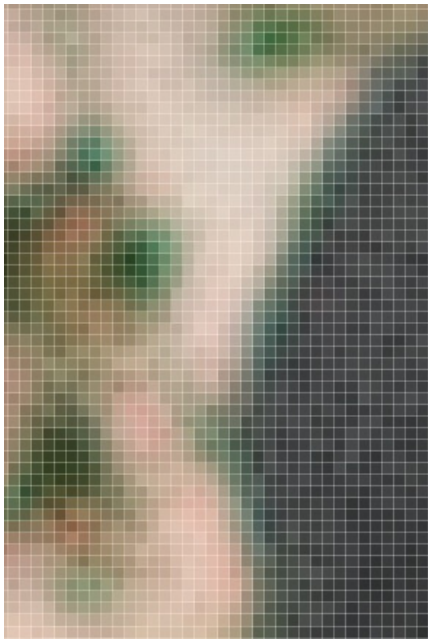
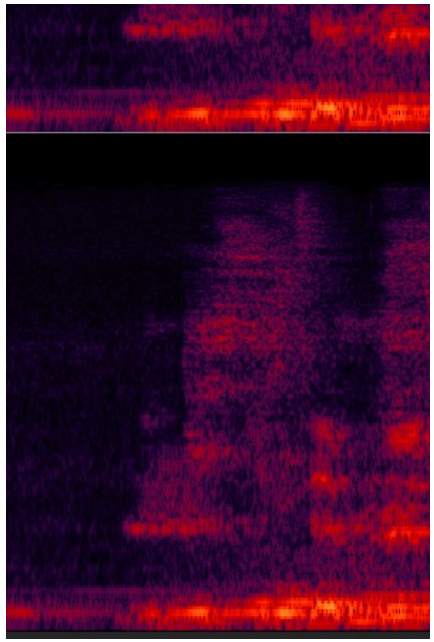
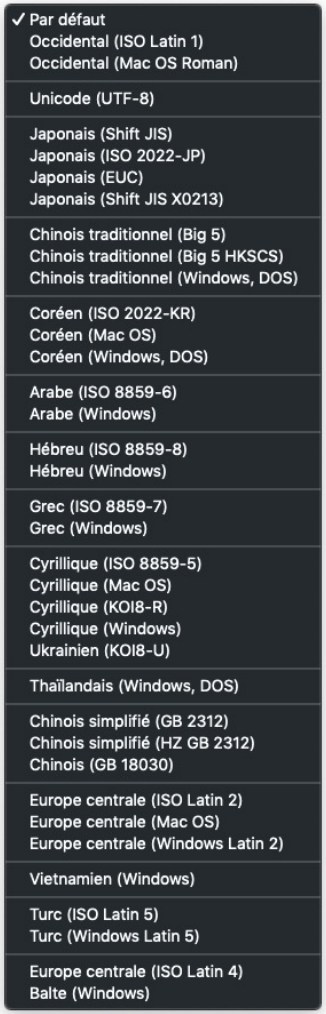


Image : des pixels



Audio : des fréquences

L’encodage des caractères



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

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

7	00A8	00A9	00AA	00AB	00AC	00A
7	00B8	00B9	00BA	00BB	00BC	00B
7	00C8	00C9	00CA	00CB	00CC	00C
7	00D8	00D9	00DA	00DB	00DC	00D
7	00E8	00E9	00EA	00EB	00EC	00E
7	00F8	00F9	00FA	00FB	00FC	00F

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

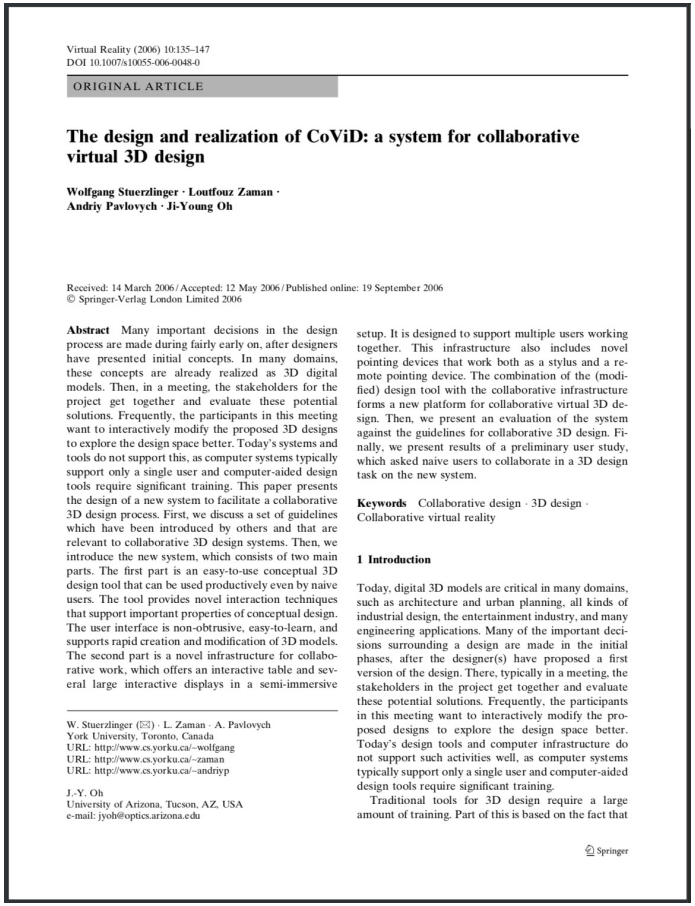
face-smiling

№	Code	Browser	Appl	Goog	FB
	U+1F600				

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

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

L’encodage des documents



Virtual Reality (2006) 10:135–147
DOI 10.1007/s10055-006-0048-0

ORIGINAL ARTICLE

The design and realization of CoViD: a system for collaborative virtual 3D design

Wolfgang Stuerzlinger · Loutfouz Zaman · Andriy Pavlovych · Ji-Young Oh

Received: 14 March 2006 / Accepted: 12 May 2006 / Published online: 19 September 2006

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 offers an interactive table and several large interactive displays in a semi-immersive 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. The combination of the (modified) design tool with the collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present an evaluation of the system against the guidelines for collaborative 3D design. Finally, we present results of a preliminary user study, which asked naive users to collaborate in a 3D design task on the new system.

Keywords Collaborative design · 3D design · Collaborative virtual reality

1 Introduction

Today, digital 3D models are critical in many domains, such as architecture and urban planning, all kinds of industrial design, the entertainment industry, and many engineering applications. Many of the important decisions surrounding a design are made in the initial phases, after the designer(s) have proposed a first version of the design. There, typically in a meeting, the stakeholders in the project get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the proposed designs to explore the design space better. Today’s design tools and computer infrastructure do not support such activities well, as computer systems typically support only a single user and computer-aided design tools require significant training.

Traditional tools for 3D design require a large amount of training. Part of this is based on the fact that

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

=

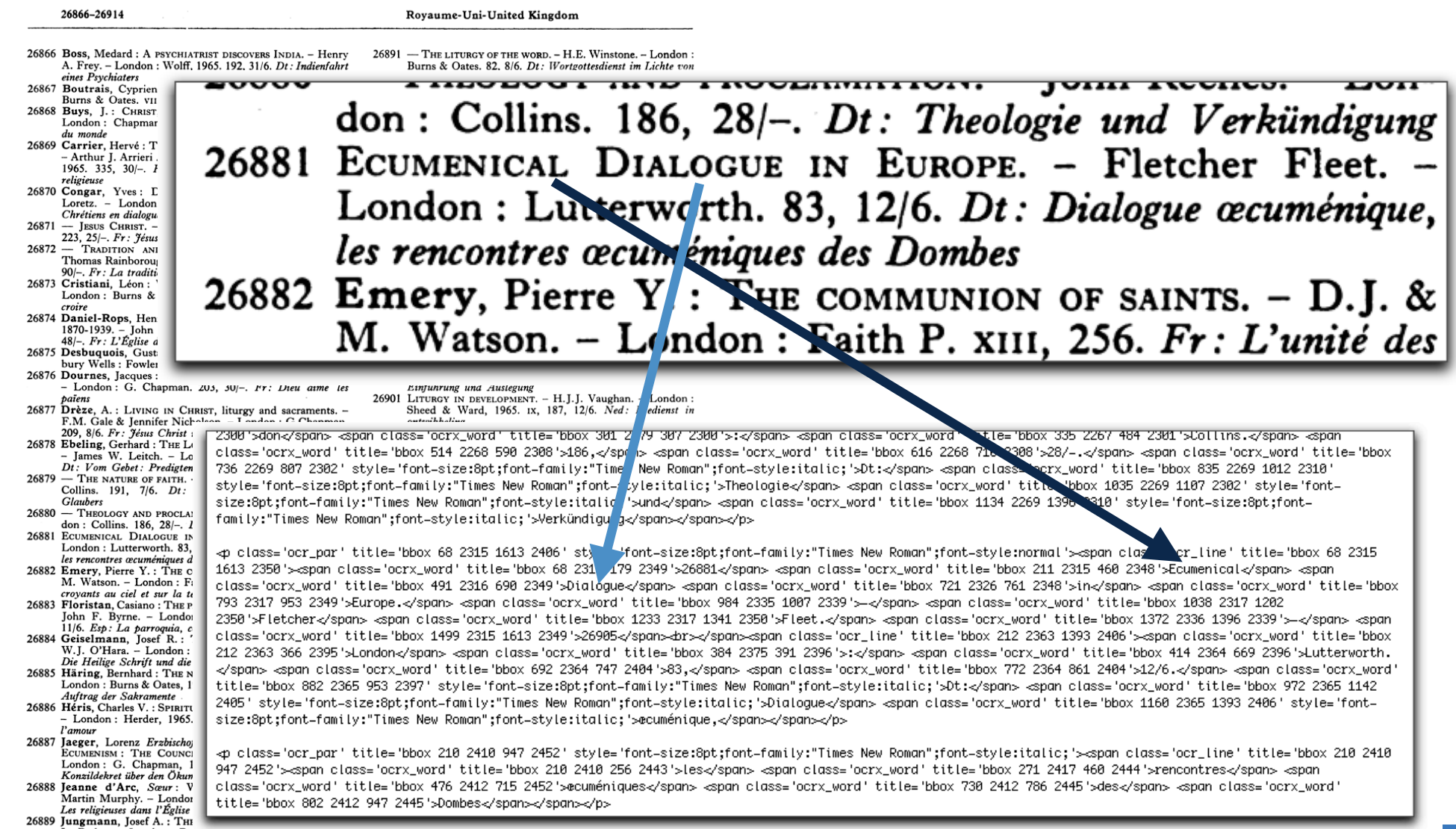
(base) [Patrice Mac-Pro-de-Patrice ~] more /Users/Patrice/Downloads/ark__67375_V
"/Users/Patrice/Downloads/ark__67375_VQC-MZJ5GK10-Q/PDF/10055_2006_Article_48.pdf"
%PDF-1.3^M%<80><84><88><8C><90><94><98><9C><A0><A4><A8><AC><B0><B4><B8><BC><C0>
<C4><C8><CC><D0><D4><D8><DC><E0><E4><E8><EC><F0><F4><F8><FC>^M1 0 obj^M<< /CropBox
nots^M[2 0 R 3 0 R 4 0 R 5 0 R 6 0 R 7 0 R 8 0 R^M] /Rotate 0 /MediaBox^M[0 0 59
0 R /Contents 10 0 R /Type /Page^M>>^Mendobj^M2 0 obj^M<< /Rect^M[367.483032 417
ype /Link /Border^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M3 0 obj^M<< /Rect^M[435.6
R18) /Subtype /Link /Border
^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M4 0 obj^M<< /Rect^M[306.142029 255.628006
k /Border^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M5 0 obj^M<< /Rect^M[430.356018 15
type /Link /Border^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M6 0 obj^M<< /Rect^M[454.
CR9) /Subtype /Link /Border^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M7 0 obj^M<< /Rec
/Dest (lCR10) /Subtype /Link /Border^M[0 0 0^M] /Type /Annot^M>>^Mendobj^M8 0 ob
.997009^M] /Dest (lCR6) /Subtype /Link /Border^M[0 0 0^M] /Type /Annot^M>>^Mendob
185 0 R^M>> /ProcSet^M[/PDF /Text^M] /ExtGState^M<< /GS1 163 0 R^M>> /ColorSpace
0 obj^M<< /Filter /FlateDecode /Length 6236^M>>^Mstream
H<89>tWM<8F><E3><B8>^Q<BD><F7><8F>^X87u<D0><ED><91>H<EA>++<B7>I^V<9B>_<97>^D X^C9
<EC>Ø<E8><wd><D1>^Q%[<9D>^?<B1>[<C9><DF>M<BD>*<CA>v0^T^L<D0>CS<FC>(V<BD><F7><AA>s
10055_2006_Article_48.pdf
%PDF-1.3^M%<80><84><88><8C><90><94><98><9C><A0><A4><A8><AC><B0><B4><B8><BC><C0>
<C4><C8><CC><D0><D4><D8><DC><E0><E4><E8><EC><F0><F4><F8><FC>^M1 0 obj^M<< /CropBox
nots^M[2 0 R 3 0
0 R /Contents 10
ype /Link /Border
R18) /Subtype /Li
/Dest (lCR5) /Sub
997009^M] /Dest (l
5029 165.997009^M
12 490.109985 165.
9 156.019012 530.
/F8 79 0 R /F2 1
183 0 R^M>>
^M>>^Mendobj^M10 0
H<89>tWM<8F><E3><
<EC>Ø<E8><wd><D1>^
<EA>_O<9F><FF>n<C
:...skipping...
%PDF-1.3^M%<80><84
8><FC>^M1 0 obj^M<<
Rotate 0 /MediaBox

The design and realization of CoViD: a system for collaborative virtual 3D design ISTEX Springer-Verlag London Springer-Verlag L
The design and realization of CoViD: a system for collaborative virtual 3D design Wolfgang Stuerzlinger York University Toronto C
CANADA Andriy Pavlovych York University Toronto CANADA Ji-Young Oh jyoh@optics.arizona.edu University of Arizona Tucso
D569A70620C3227A0BF16EF149FE0E7980229167 ark:/67375/VQC-MZJ5GK10-Q s10055-006-0048-0 10.1007/s10055-006-0048
4338 1434-9957 Springer-Verlag London 2006 10 2 135 147 pub2TEI-ISTEX A set of style sheets for converting XML documents er
common TEI format. We use TEI Abstract Many important decisions in the design process are made during fairly early on, after desig
domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get together an
participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today’s systems
typically support only a single user and computer-aided design tools require significant training. This paper presents the design of a n
process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D design sy
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 user
support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rapid creation and
infrastructure for collaborative work, which offers an interactive table and several large interactive displays in a semi-immersive setu
together. This infrastructure also includes novel pointing devices that work both as a stylus and a remote pointing device. The combin
collaborative infrastructure forms a new platform for collaborative virtual 3D design. Then, we present an evaluation of the system ag
Finally, we present results of a preliminary user study, which asked naive users to collaborate in a 3D design task on the new system.
Vision Artificial Intelligence (incl. Robotics) Computing Methodologies Computer Graphics Collaborative design 3D design Collabo
Accepted ePublished Virtual Reality (2006) 10:135–147 DOI 10.1007/s10055-006-0048-0 O R I G I N A L A R T I C L E The design
virtual 3D design Wolfgang Stuerzlinger Æ Loutfouz Zaman Æ Andriy Pavlovych Æ Ji-Young Oh Received: 14 March 2006 / Accep
2006 Springer-Verlag London Limited 2006 Abstract Many important decisions in the design process are made during fairly early on,
many domains, these concepts are already realized as 3D digital models. Then, in a meeting, the stakeholders for the project get togetl
the participants in this meeting want to interactively modify the proposed 3D designs to explore the design space better. Today’s syste
systems typically support only a single user and computer-aided design tools require significant training. This paper presents the desig
design process. First, we discuss a set of guidelines which have been introduced by others and that are relevant to collaborative 3D de
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 nai
techniques that support important properties of conceptual design. The user interface is non-obtrusive, easy-to-learn, and supports rap
second part is a novel infrastructure for collaborative work, which offers an interactive table and several large interactive displays in a
A. Pavlovych York University, Toronto, Canada URL: http://www.cs.yorku.ca/~wolfgang URL: http://www.cs.yorku.ca/~zaman URL
University of Arizona, Tucson, AZ, USA e-mail: jyoh@optics.arizona.edu setup. It is designed to support multiple users working tog
devices that work both as a stylus and a remote pointing device. The combination of the (modified) design tool with the collaborative
virtual 3D design. Then, we present an evaluation of the system against the guidelines for collaborative 3D design. Finally, we presen
naive users to collaborate in a 3D design task on the new system. Keywords Collaborative design 3D design Collaborative virtual real
critical in many domains, such as architecture and urban planning, all kinds of industrial design, the entertainment industry, and many
decisions surrounding a design are made in the initial phases, after the designer(s) have proposed a first version of the design. There, t
get together and evaluate these potential solutions. Frequently, the participants in this meeting want to interactively modify the propos

PDF to text

Evitons les PDF…

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. Arrieri . 1965. 335, 30/–. *religieuse*

26870 **Congar**, Yves : L Loretz. – London *Chrétiens en dialogu*

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'Église a*

26875 **Desbuquois**, Gust bury Wells : Fowler

26876 **Dournes**, Jacques : – London : G. Chapman. 203, 30/–. *Fr : Dieu aime les païens*

26877 **Drèze**, A. : LIVING IN CHRIST, liturgy and sacraments. – F.M. Gale & Jennifer Nich – 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 PROCLA 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 *croyants au ciel et sur la t*

26883 **Floristan**, Casiano : THE P John F. Byrne. – Londo 11/6. *Esp : La parroquia, c*

26884 **Geiselmann**, Josef R. : W.J. O'Hara. – London : *Die Heilige Schrift und die*

26885 **Häring**, Bernhard : THE N London : Burns & Oates, 1 *Auftrag der Sakramente*

26886 **Héris**, Charles V. : SPIRITU – London : Herder, 1965. *l'amour*

26887 **Jaeger**, Lorenz *Erzbischo* ECUMENISM : THE COUNCI London : G. Chapman, 1 *Konzildekret über den Ökun*

26888 **Jeanne d'Arc**, *Sœur* : V Martin Murphy. – Londo *Les religieuses dans l'Église*

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*

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*

Einführung und Auslegung

26901 LITURGY IN DEVELOPMENT. – H.J.J. Vaughan. – London : Sheed & Ward, 1965. IX, 187, 12/6. *Ned : Liedienst in*

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

26881 Ecumenical Dialogue in Europe. – Fletcher Fleet. – 26905
London : Lutterworth. 83, 12/6. Dt: Dialogue œcuménique,

les rencontres œcuméniques des Dombes

26914 Roman Catholic Church. *2d Vatican Ecumenical Council.* DECLARATION ON RELIGIOUS LIBERTY. – Thomas Atthill. – London : Catholic Truth Society. 19, 1/–. *Lat : De libertate religiosa*

567

Les fleurs du mal / par Charles Baudelaire

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

.MJ LECTEUR

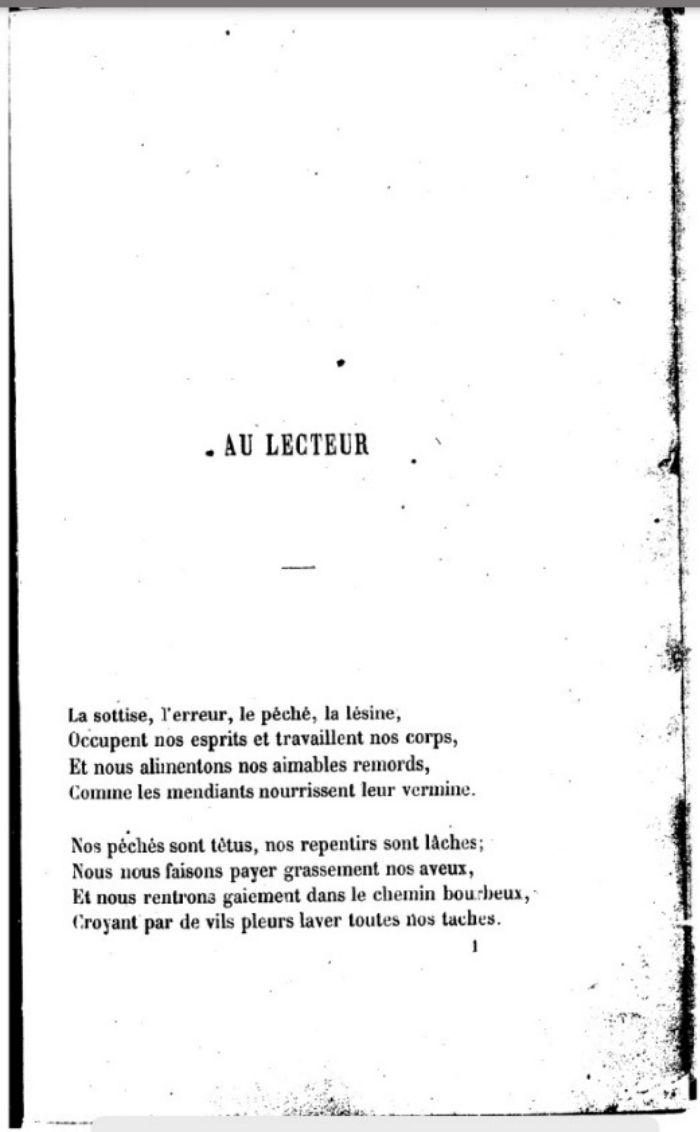
La sottise, l'erreur, le péché, la lésine,

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 w
Croyant par de vils pleurs laver toutes nos taches.
i

1



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

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

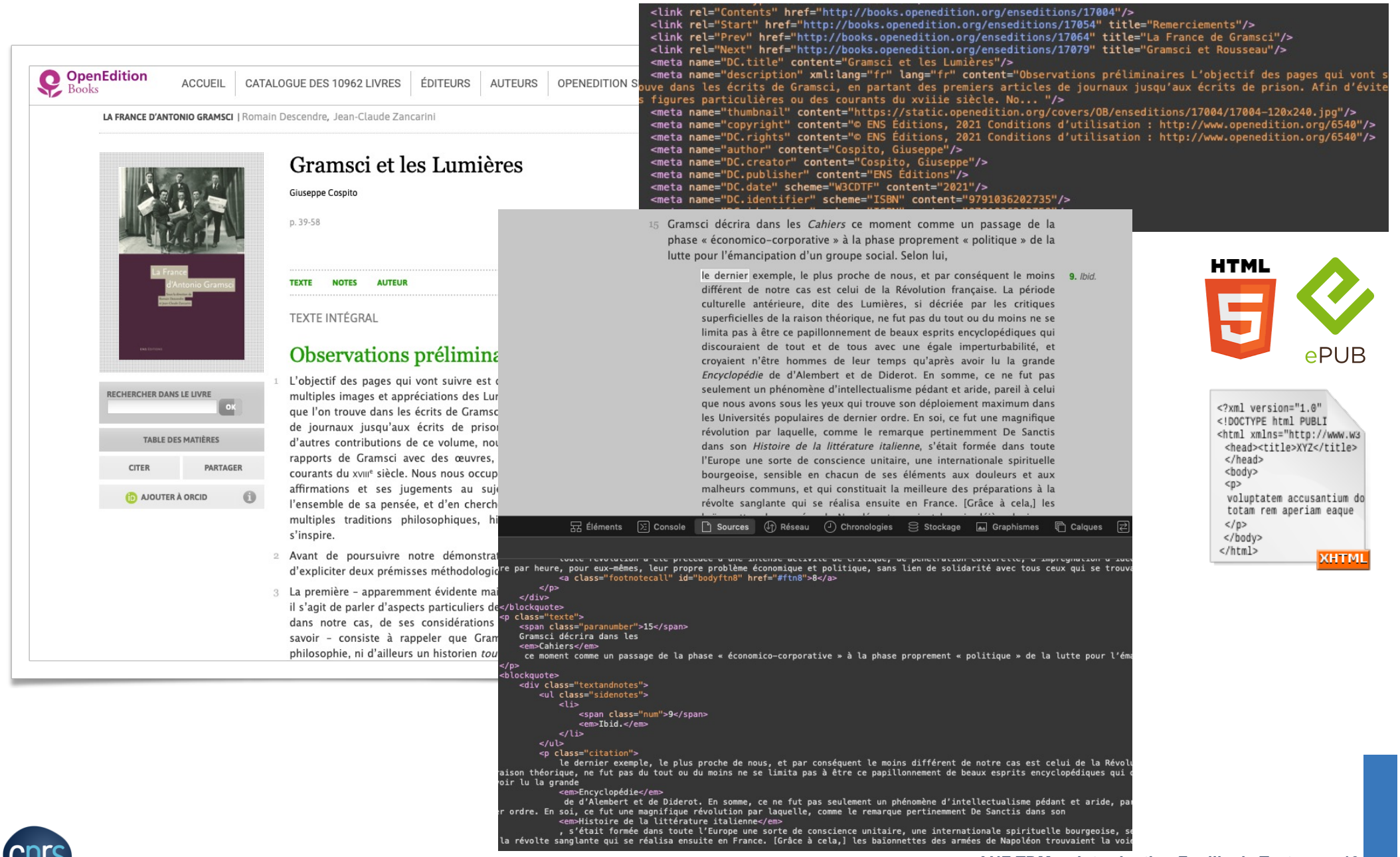
XML

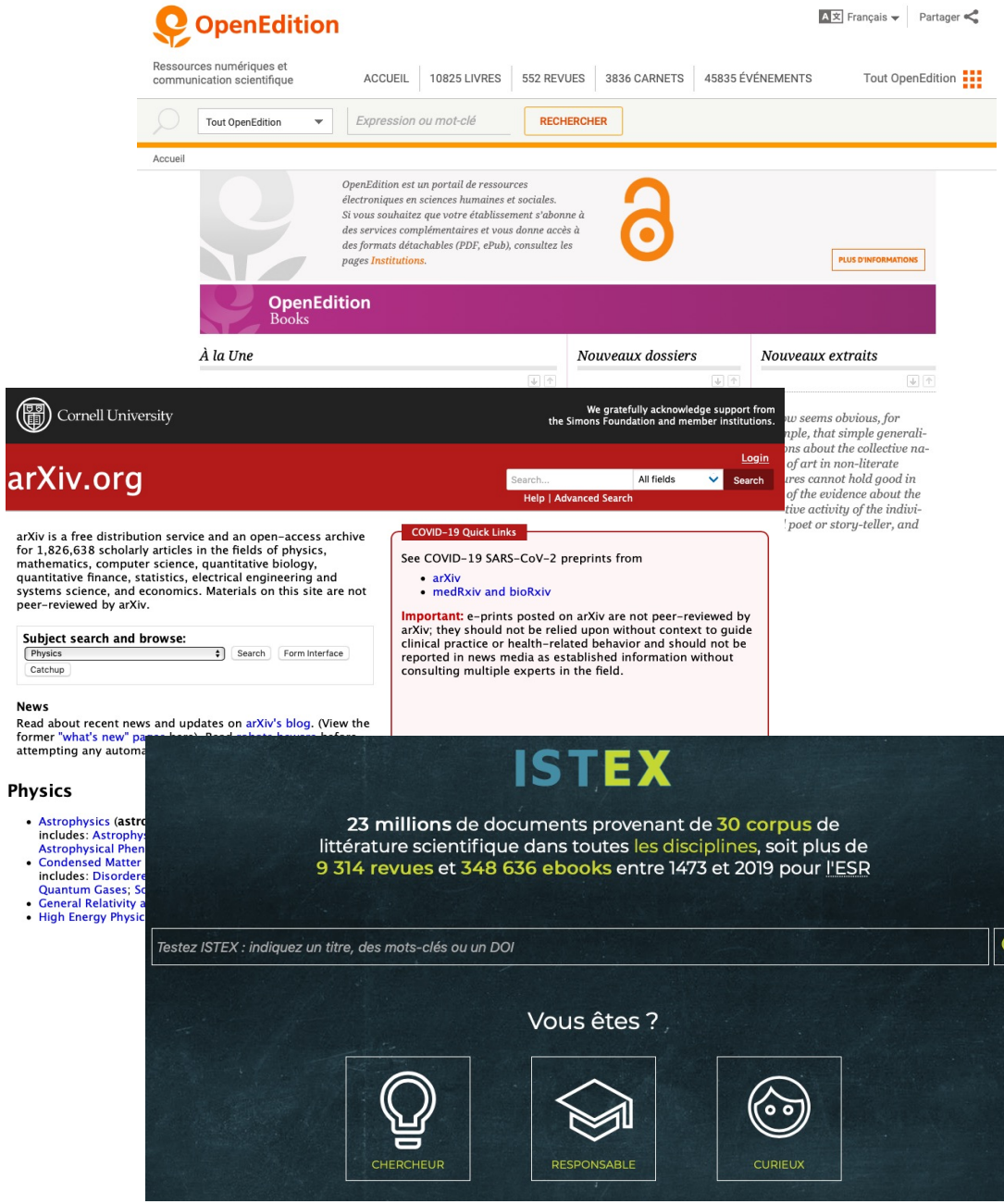
```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:ns1="http://standoff.proposal"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:noNamespaceSchemaLocation="https://xml-schema.delivery.istex.fr/formats/tei-istex.xsd">
   <teiHeader>
      <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>
```

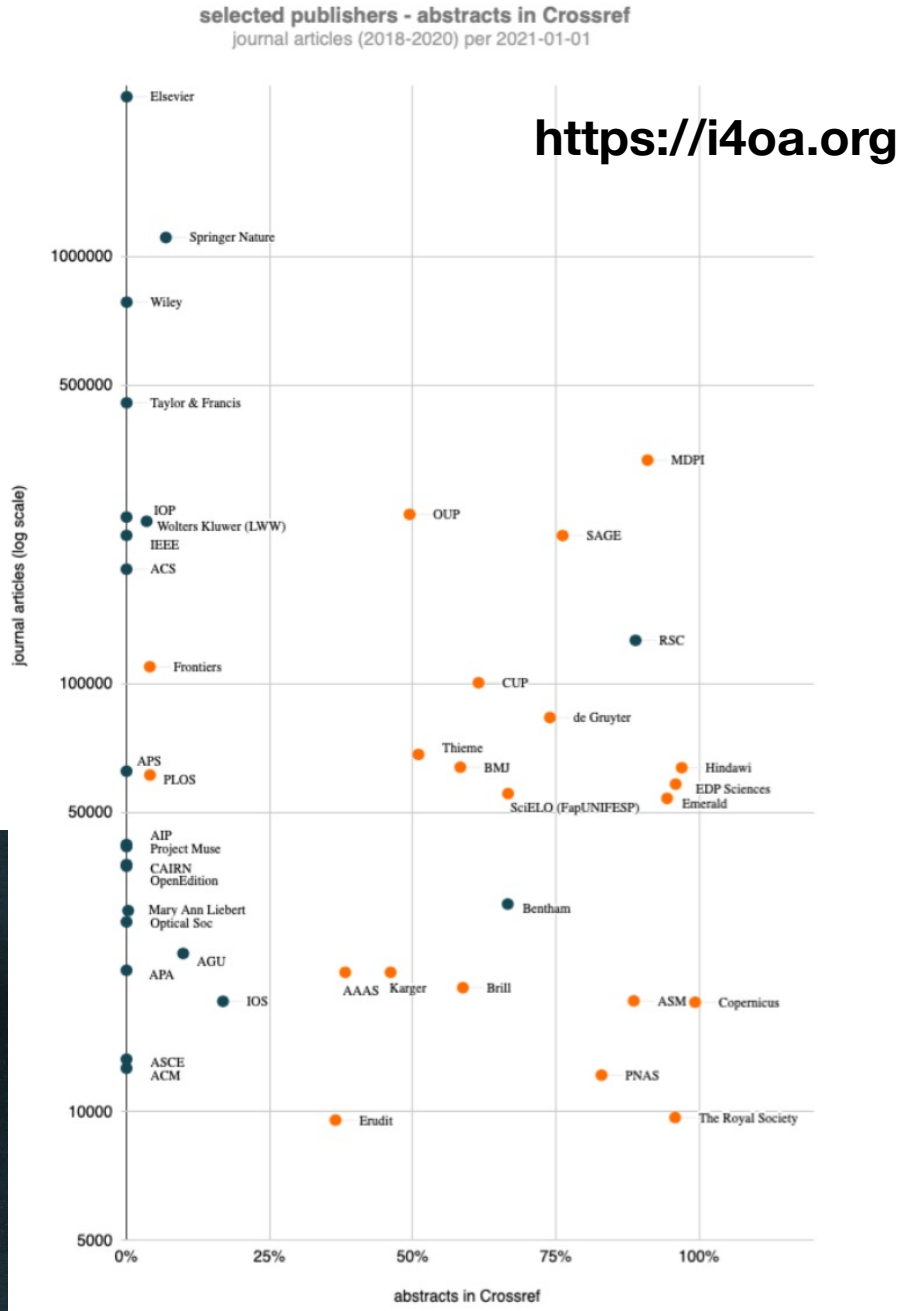
```
Abstract Many important decisions in the design
process are made during fairly early on, after designers
have presented initial concepts. In many domains,
these concepts are already realized as 3D digital
models. Then, in a meeting, the stakeholders for the
project get together and evaluate these 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 (&amp;amp;)  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







Texte intégral

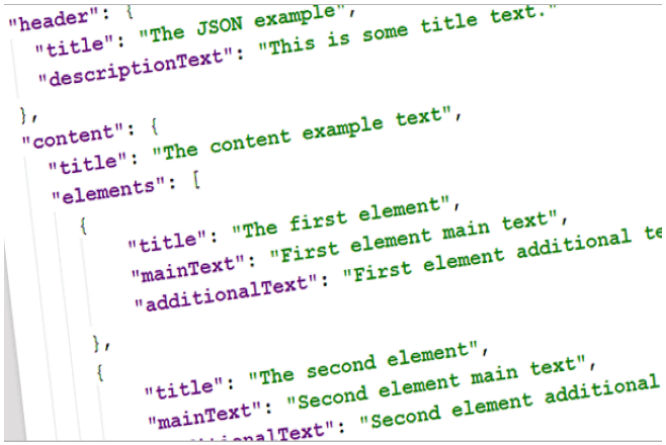
Résumés des articles

D’autres formats ouverts

```
"jour";"nomReg";"numReg";"incid_rea"
2020-03-19;"Auvergne-RhÛne-Alpes";84;44
2020-03-19;"Bourgogne-Franche-ComtÈ";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 RÈunion";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-CÙte d'Azur";93;25
2020-03-20;"Auvergne-RhÛne-Alpes";84;16
2020-03-20;"Bourgogne-Franche-ComtÈ";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



JSON

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

```yaml
---
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 × constituency parse × dependency parse ×

— Language —

French

Submit

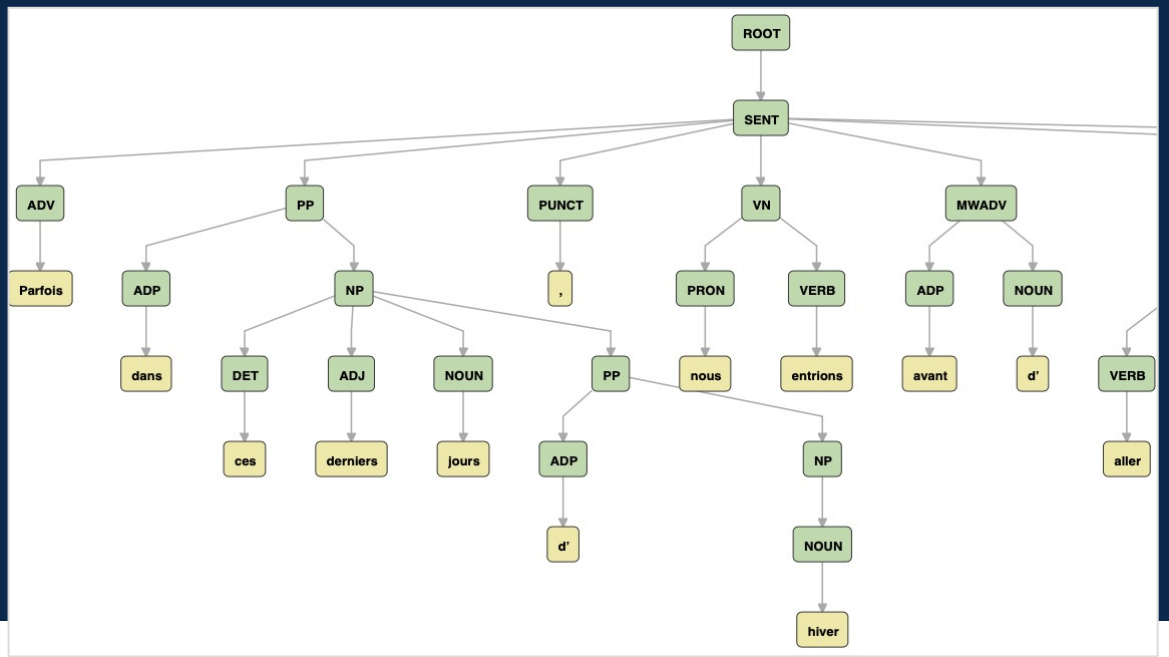
Part-of-Speech:

	ADV	ADP	DET	ADJ	NOUN	ADP	NOUN	PUNCT	PRON	VERB	ADP	NOUN	VERB	PRON	VERB	ADP	PRON	ADP	DET	ADJ	NOUN	PRON	PRON	VERB	ADV	CCONJ	PRON	PROPN	PUNCT		
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	,			
	NOUN	ADP	NOUN	PUNCT	AUX	VERB	ADP	DET	ADJ	NOUN	ADP	DET	NOUN	ADP	NOUN	ADP	PRON	PRON	VERB	NOUN	PUNCT											
	collectionneur	de	marque	,	était	salué	avec	une	particulière	déférence	par	les	marchands	de	tableaux	chez	qui	elles	avaient	lieu	.											

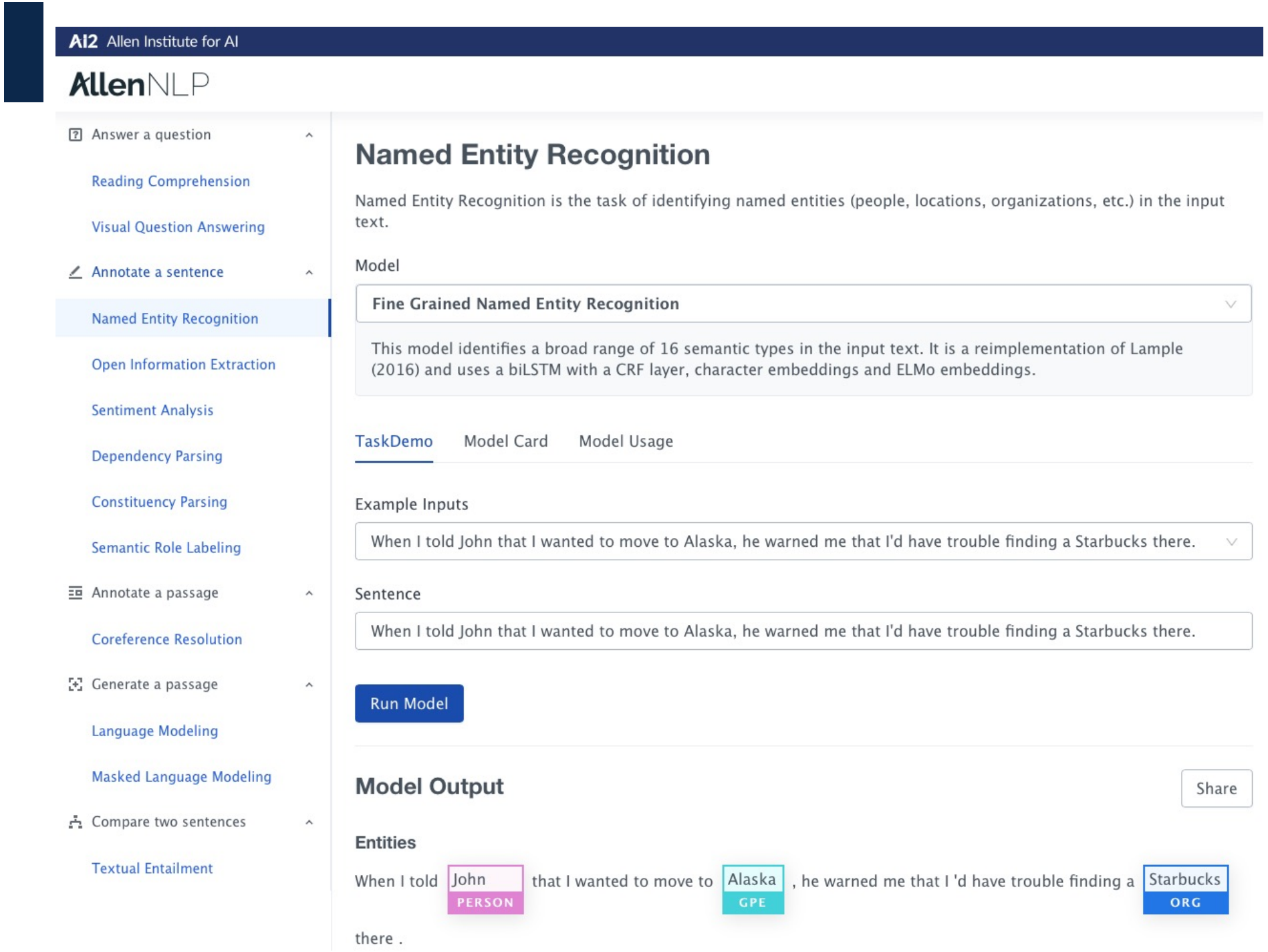
Dépendances

Unités lexicales et parties du discours

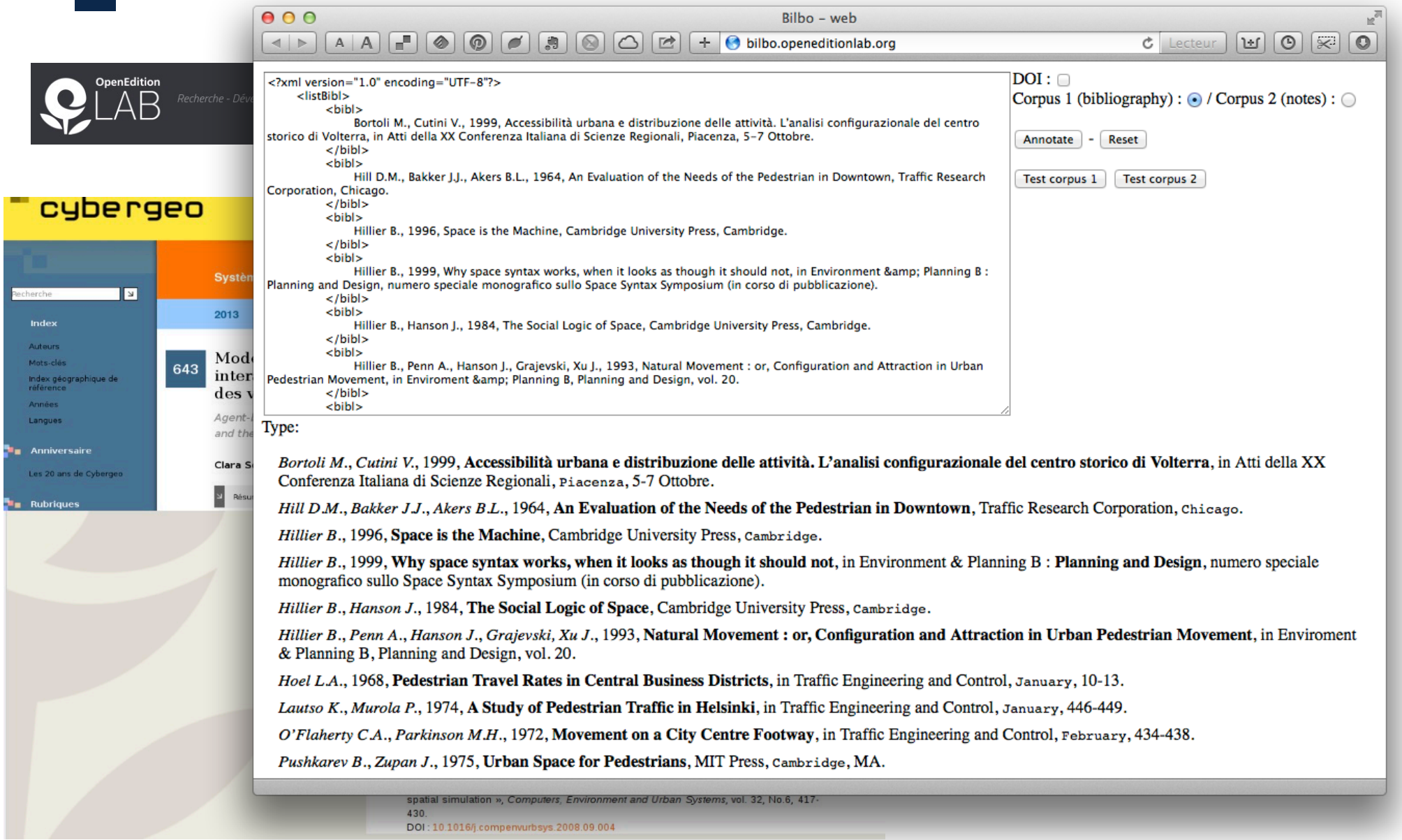
Constituants



© Manish Para
https://medium.com/datadriveninvestor/ai-pos-parts-of-speech-tagger-part-2-fcf55f891d10


AI2 Allen Institute for AI
AllenNLP
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 PERSON that I wanted to move to Alaska GPE , he warned me that I 'd have trouble finding a Starbucks ORG there .

Enrichir par la création de liens entre documents



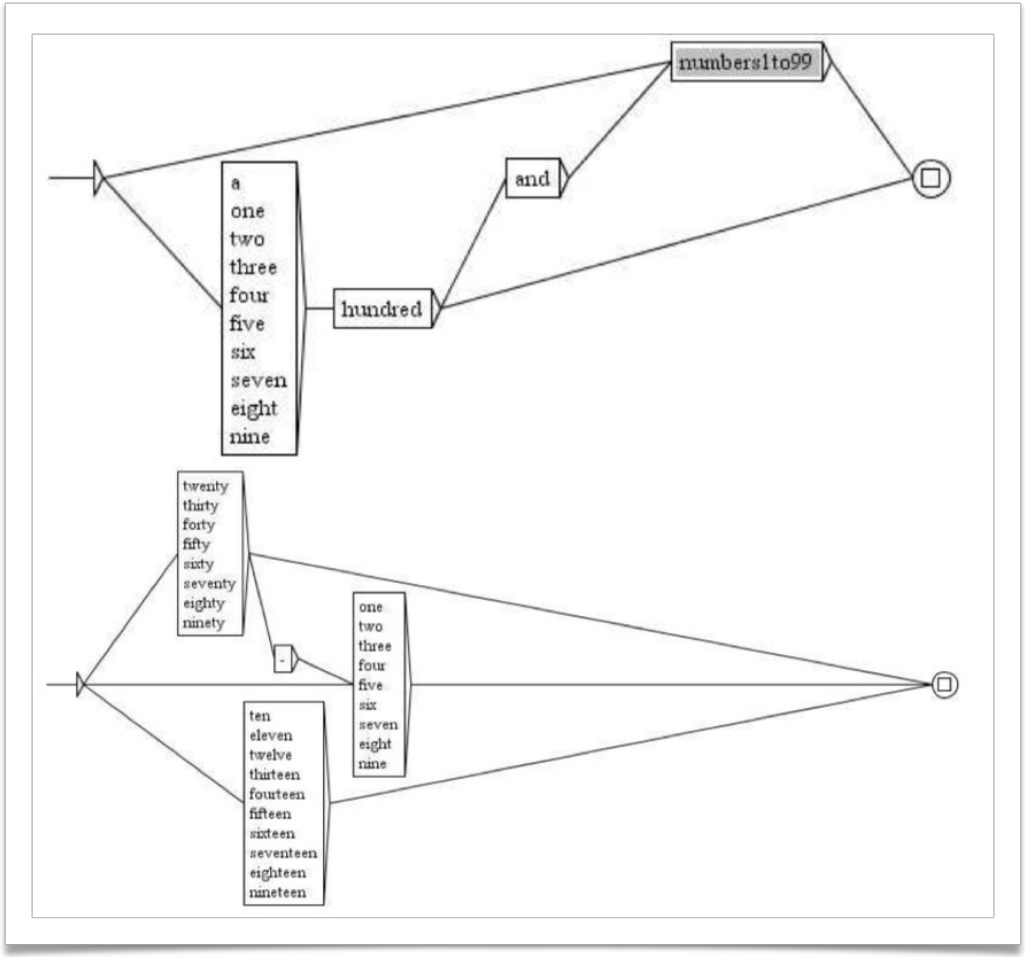
Test : http://bilbo.openeditionlab.org

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

Des approches informatiques et des ressources linguistiques

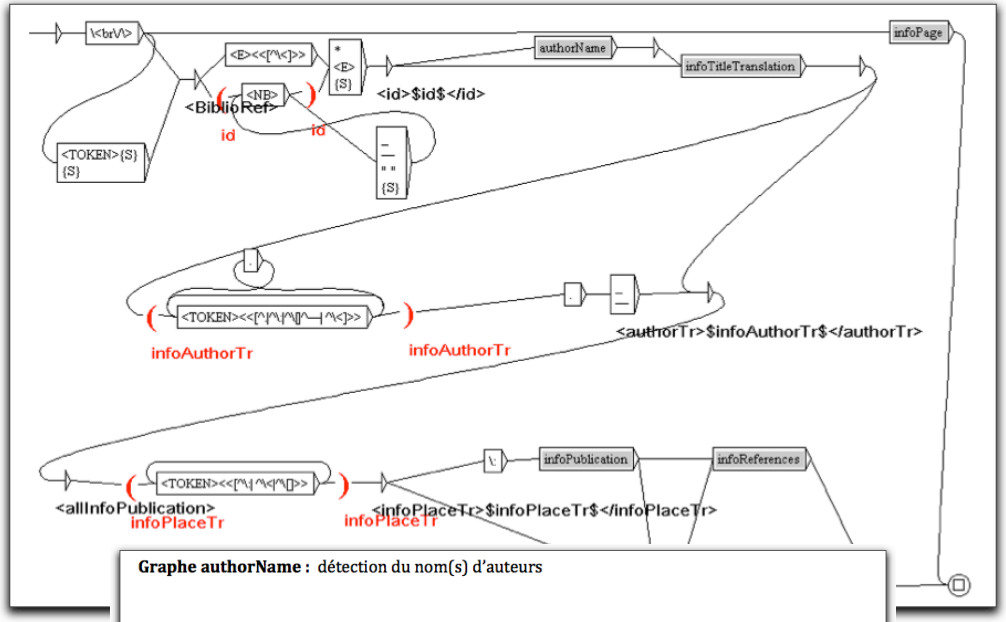
04.06.19

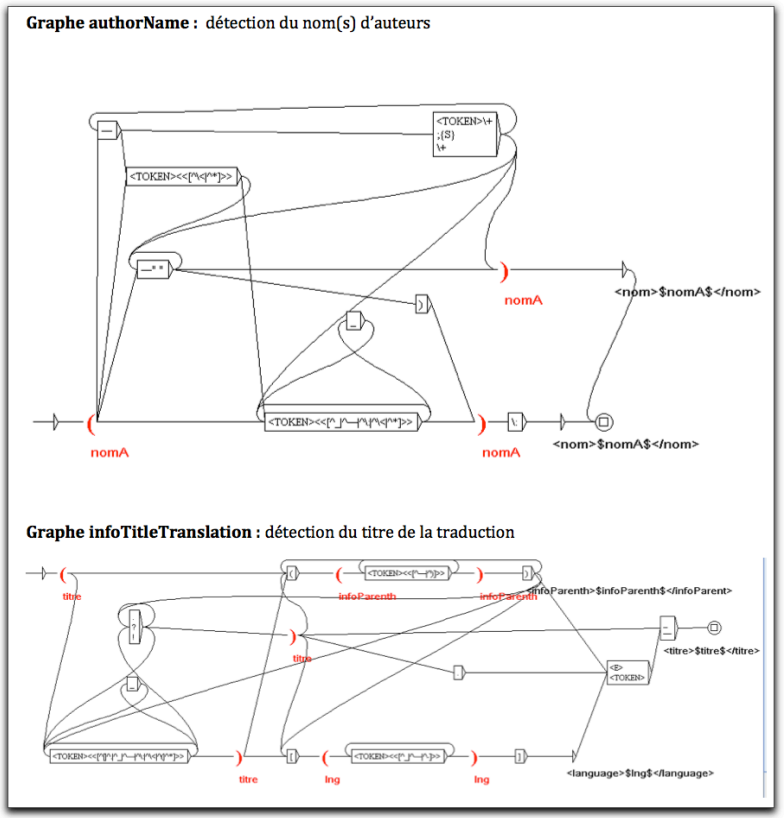
Des exemples d'automates



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

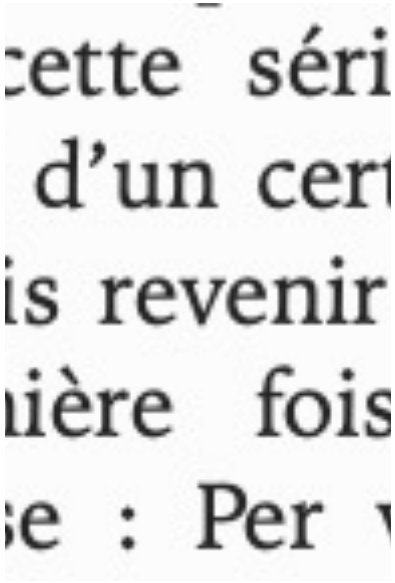
https://hal.archives-ouvertes.fr/hal-00145253/document





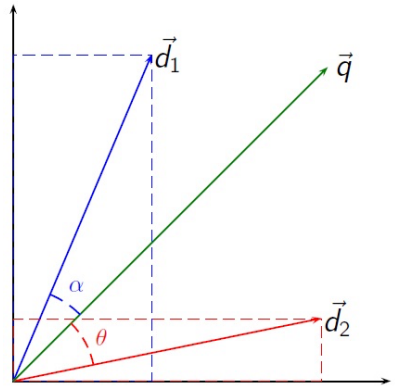
Etude pour l'UNESCO (Univ. Avignon, Open Edition 2011)

Des formes, des distributions et des vecteurs



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

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



Des vecteurs

$$\text{Similarité}(d_1, d_2) \approx \overrightarrow{d_1}.\overrightarrow{d_2}$$

$$\text{Similarité}(d_1, d_2) \approx \cos(\overrightarrow{d_1}, \overrightarrow{d_2})$$

Mot	Probabilité

Des modèles de langues

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

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

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.

Google Books

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 i 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
```

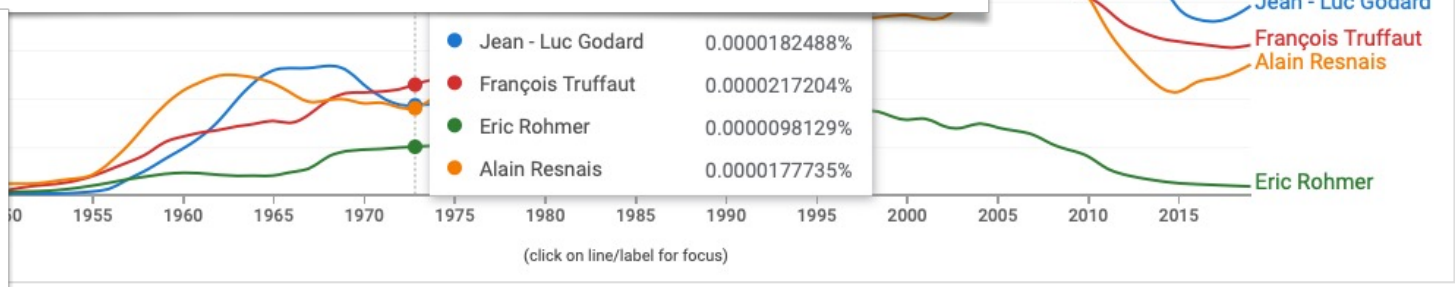
French

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

The doctor ran to the emergency room to see the

Sentence

The doctor ran to the emergency room to see the

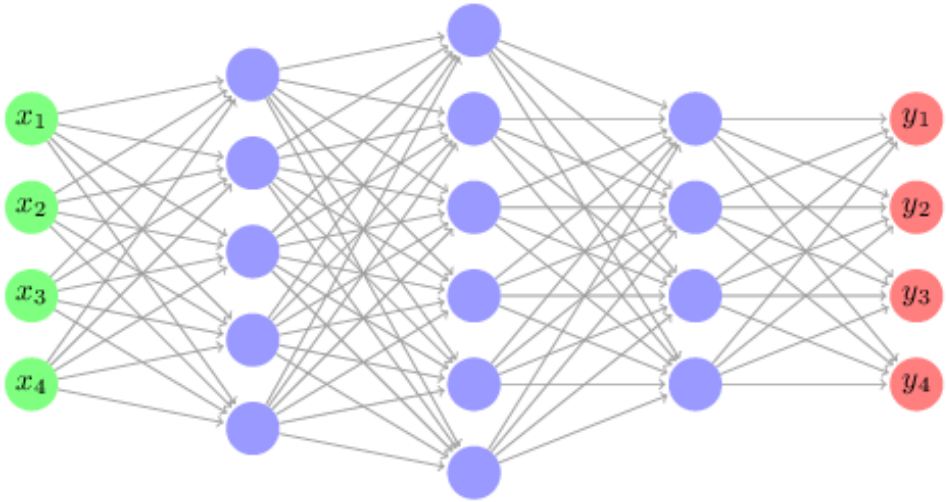
Run Model

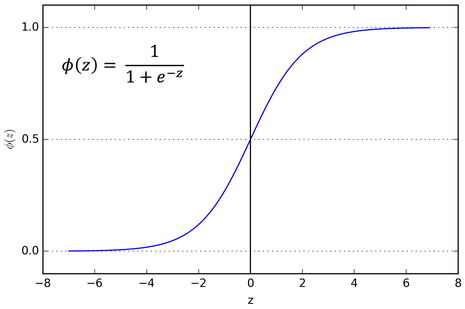
Model Output

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 %

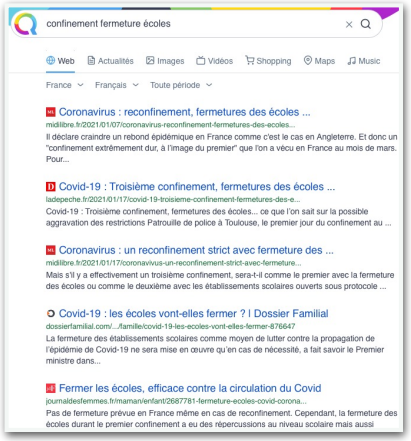
Approches numériques : des scores



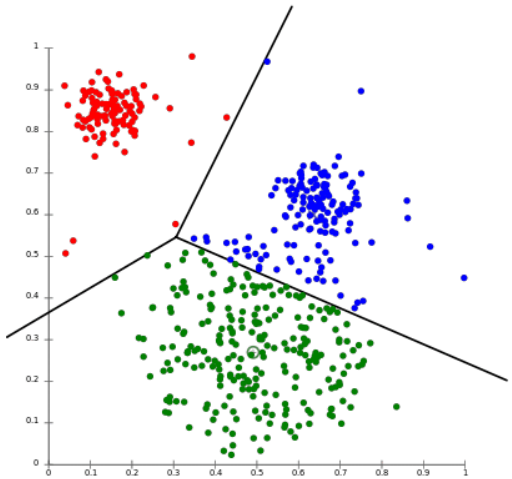


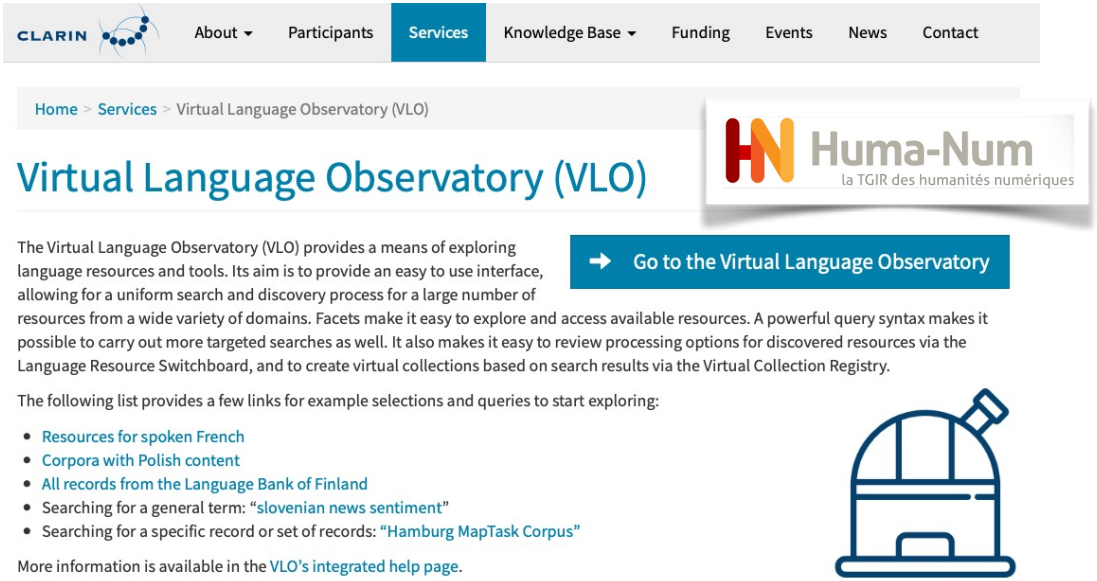
Data

Ranking / Scores



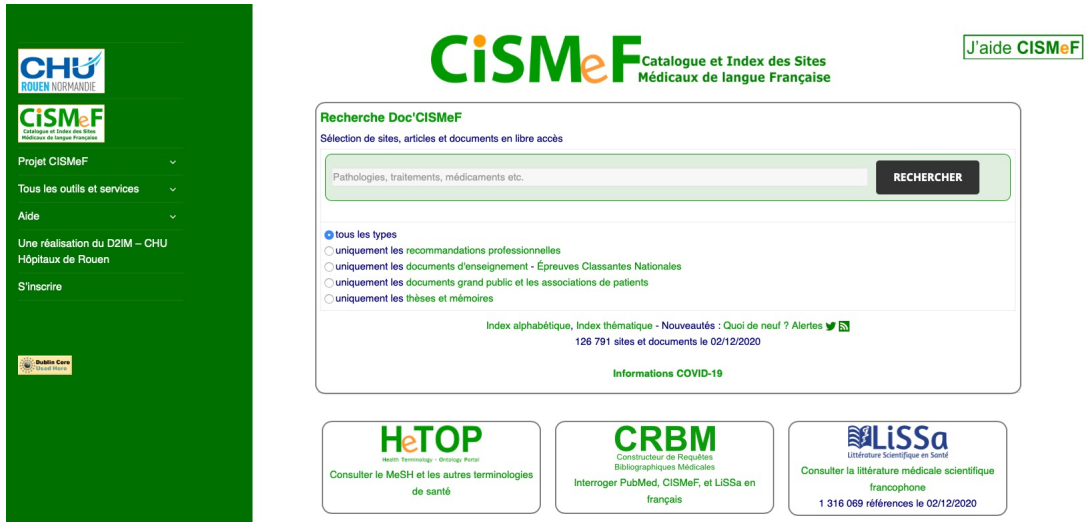
Décision ?



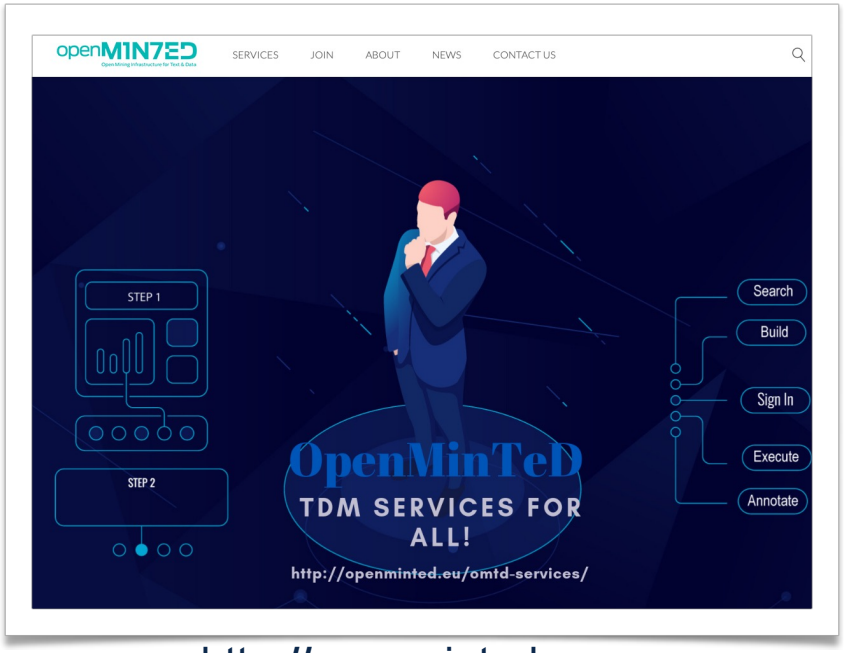


https://www.clarin.eu/content/virtual-language-observatory-vlo

https://www.ortolang.fr



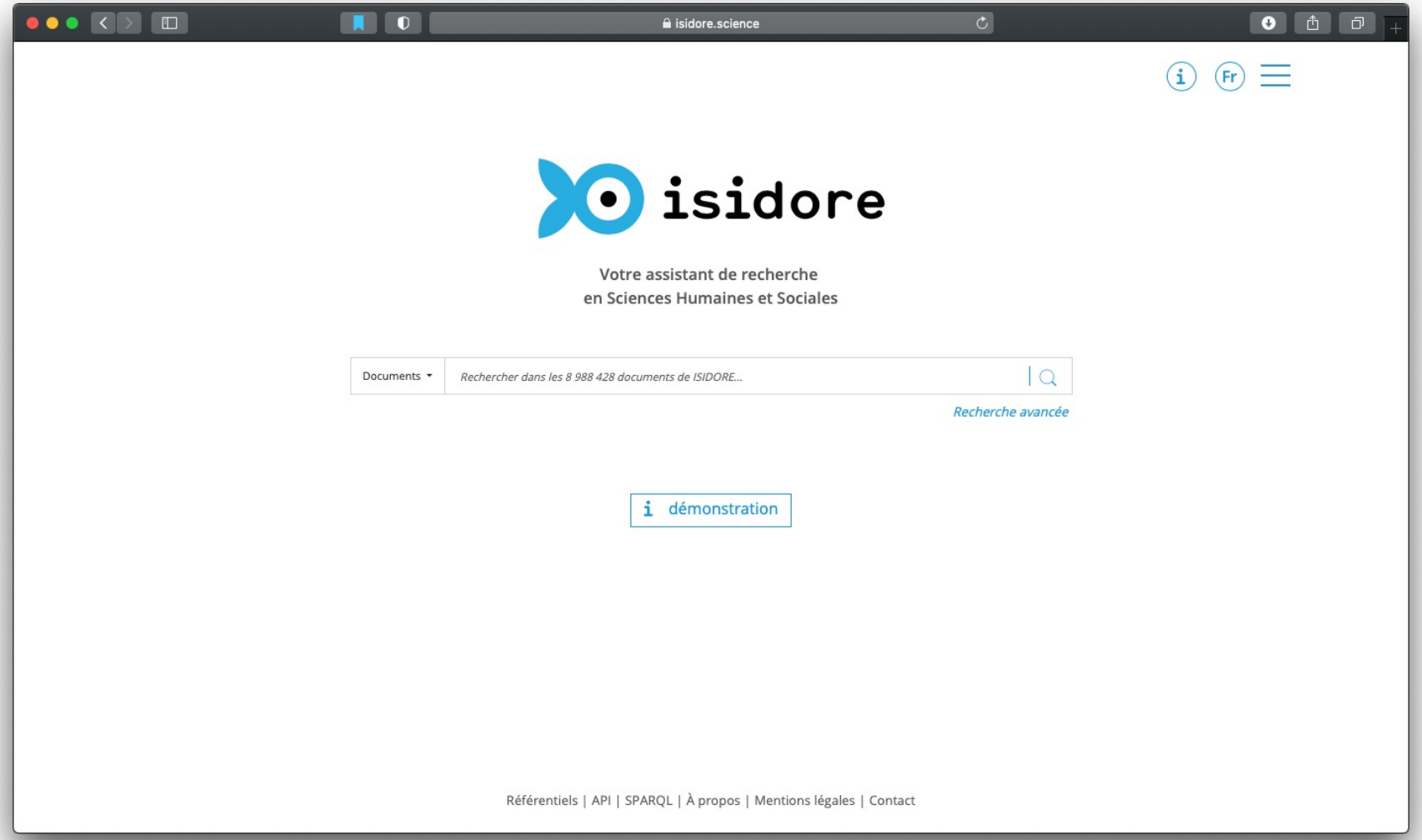
http://www.chu-rouen.fr/cismef/



http://openminted.eu

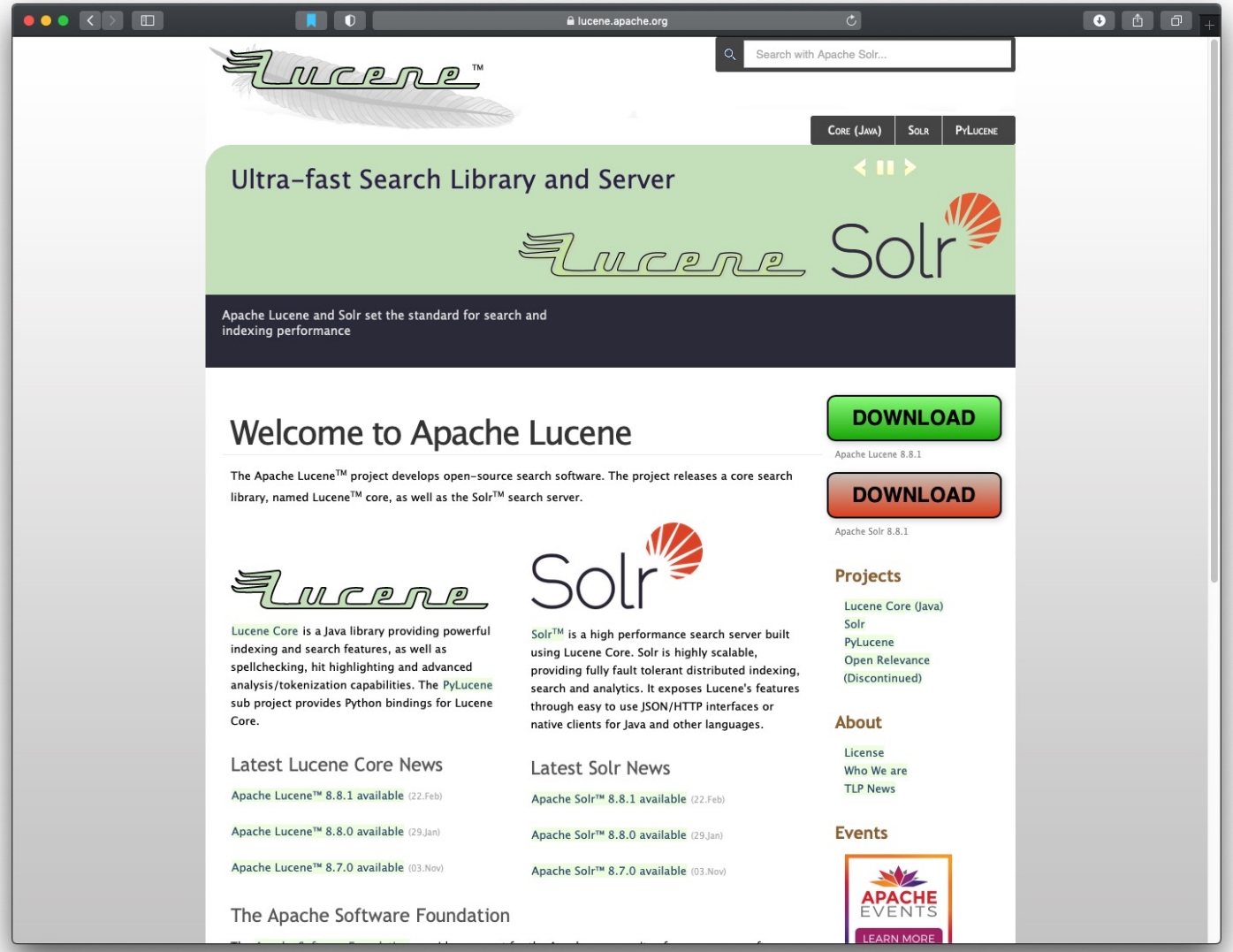
Des environnements logiciels pour le développement et l'expérimentation

Isidore : un moteur pour les SHS

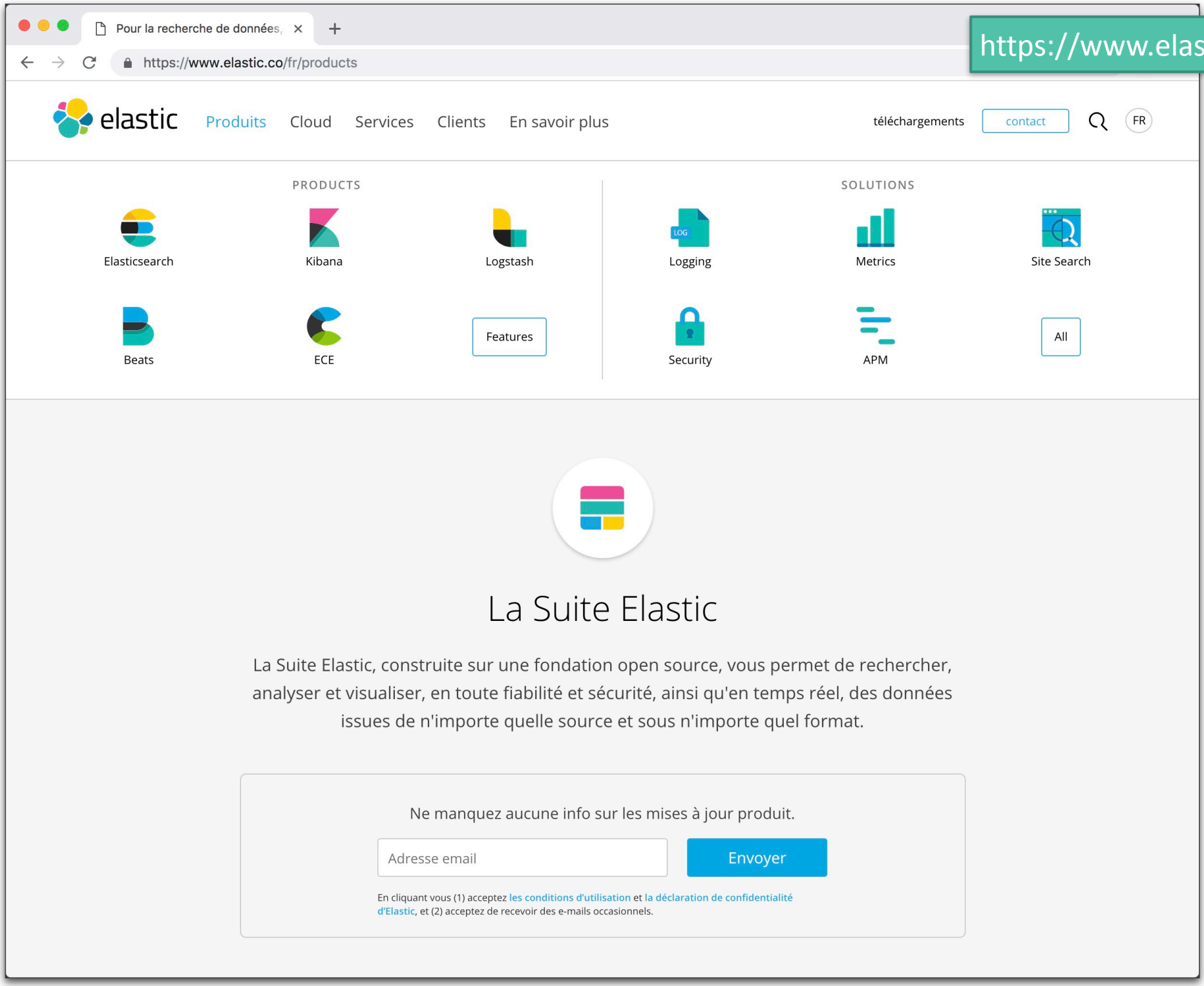


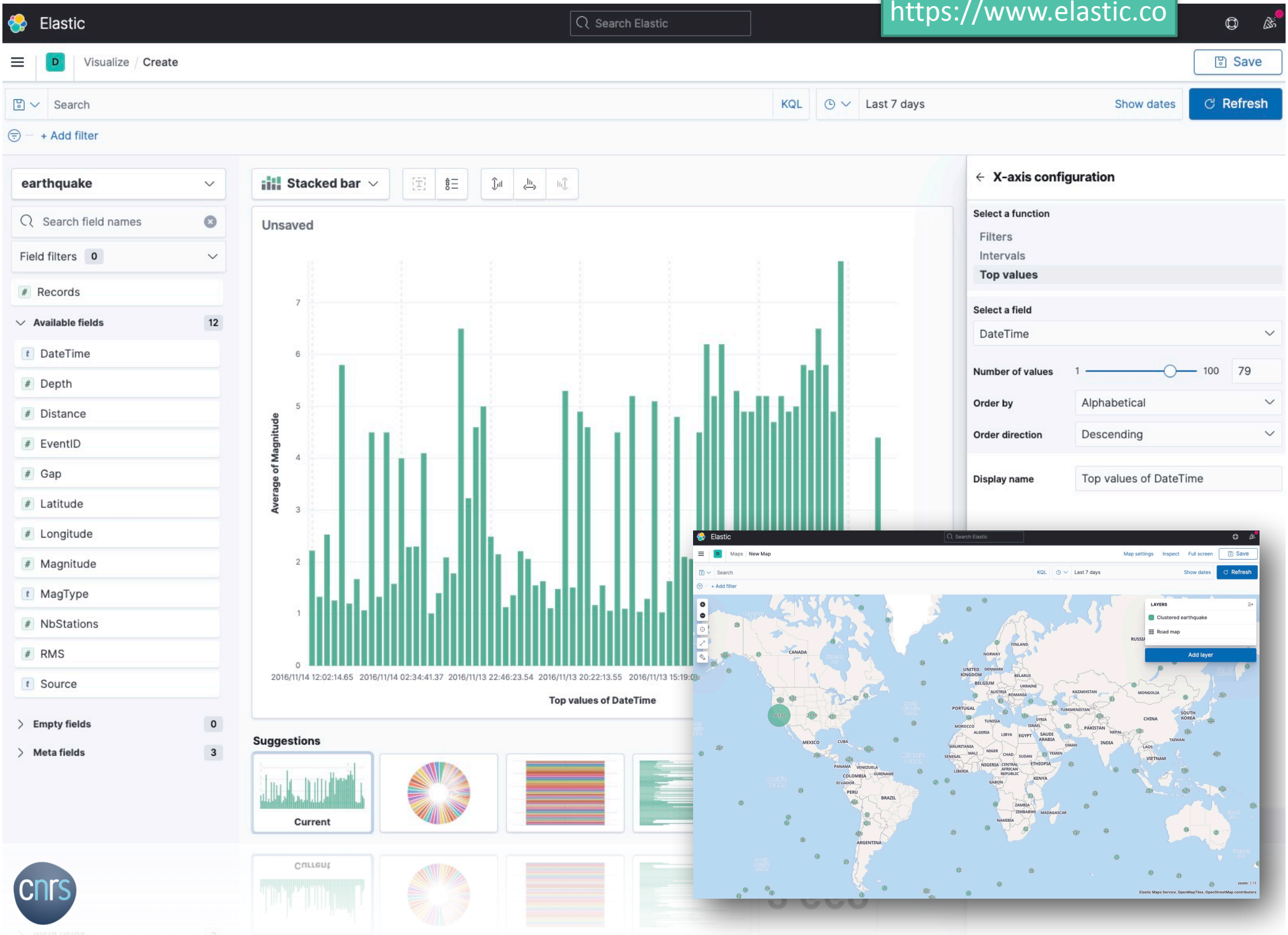
https://isidore.science

Apache Lucene : indexation et recherche d'information



https://lucene.apache.org


Pour la recherche de données,
https://www.elastic.co/fr/products
https://www.elastic.co
elastic
Produits
Cloud
Services
Clients
En savoir plus
téléchargements
contact
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.


Elastic
Search Elastic
https://www.elastic.co
Visualize / Create
Save
Search
KQL
Last 7 days
Show dates
Refresh
+ Add filter
earthquake
Search field names
Field filters 0
Records
Available fields 12
DateTime
Depth
Distance
EventID
Gap
Latitude
Longitude
Magnitude
MagType
NbStations
RMS
Source
Empty fields 0
Meta fields 3
Stacked bar
Unsaved
Average of Magnitude
Top values of DateTime
X-axis configuration
Select a function
Filters
Intervals
Top values
Select a field
DateTime
Number of values 1 100 79
Order by Alphabetical
Order direction Descending
Display name Top values of DateTime
Suggestions
Current
Maps New Map
Map settings
Inspect
Full screen
Save
Refresh
LAYERS
Clustered earthquake
Road map
Add layer
cnrs

https://www.cs.waikato.ac.nz/ml/weka/

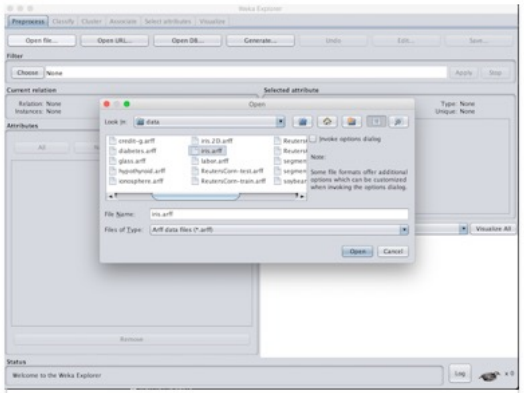
WEKA

The workbench for machine learning

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

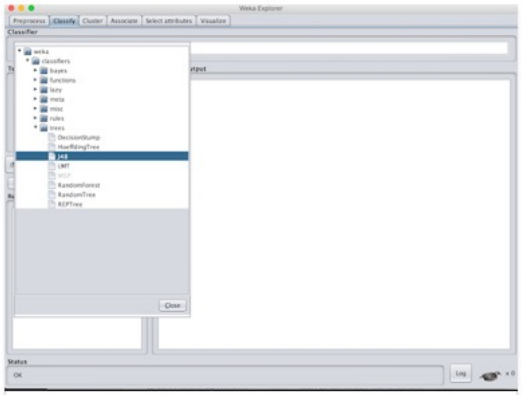
Machine Learning without Programming

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



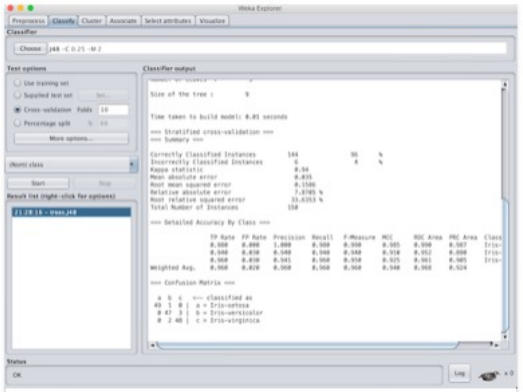
Open a dataset

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/

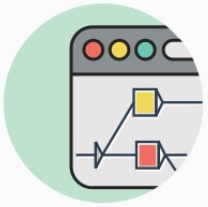
Unitex/GramLab est une suite logicielle libre, multiplateforme, multilingue, fondée sur des dictionnaires et des grammaires pour l'analyse de corpus

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

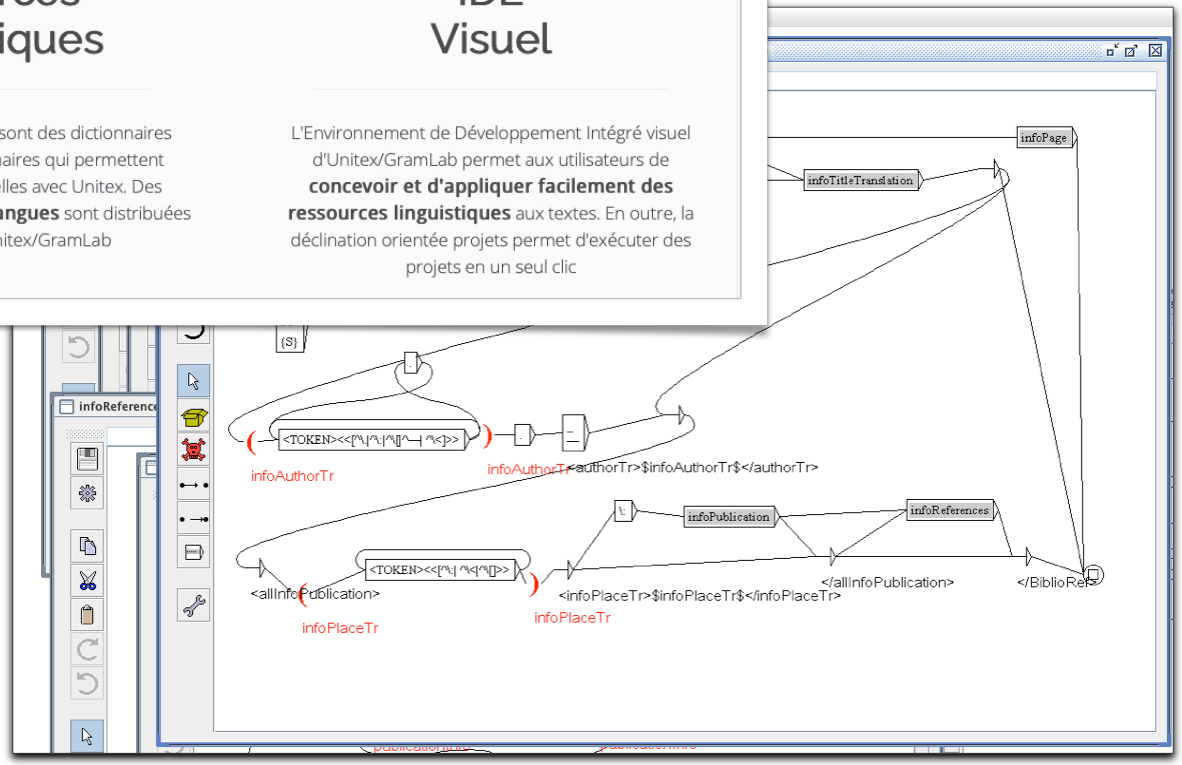
Les ressources linguistiques sont des dictionnaires électroniques et des grammaires qui permettent l'analyse de données textuelles avec Unitex. Des ressources pour **plus de 22 langues** sont distribuées actuellement avec Unitex/GramLab



IDE Visuel

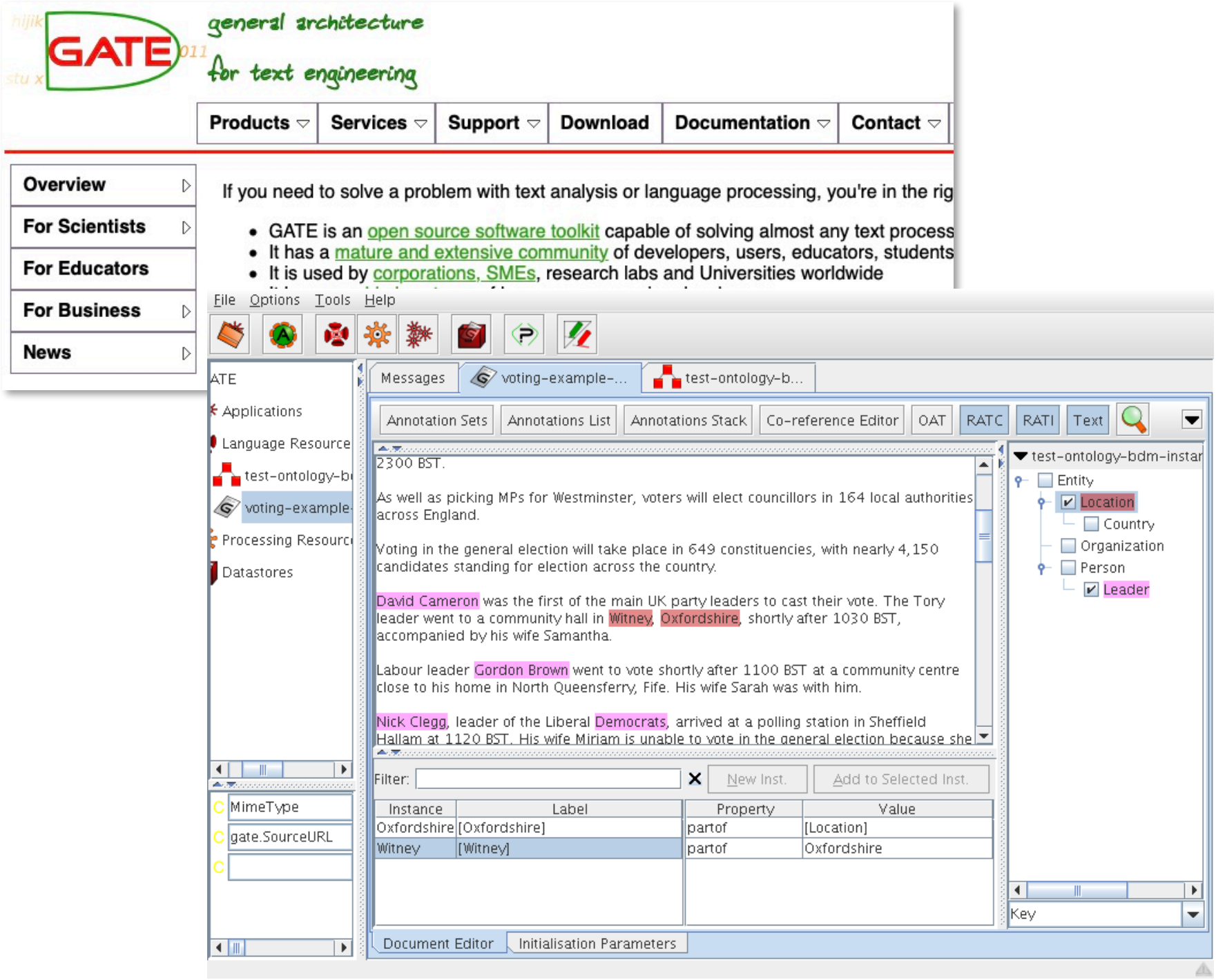
L'Environnement de Développement Intégré visuel d'Unitex/GramLab permet aux utilisateurs de **concevoir et d'appliquer facilement des ressources linguistiques** aux textes. En outre, la déclination orientée projets permet d'exécuter des projets en un seul clic

règles syntaxiques ou sémantiques

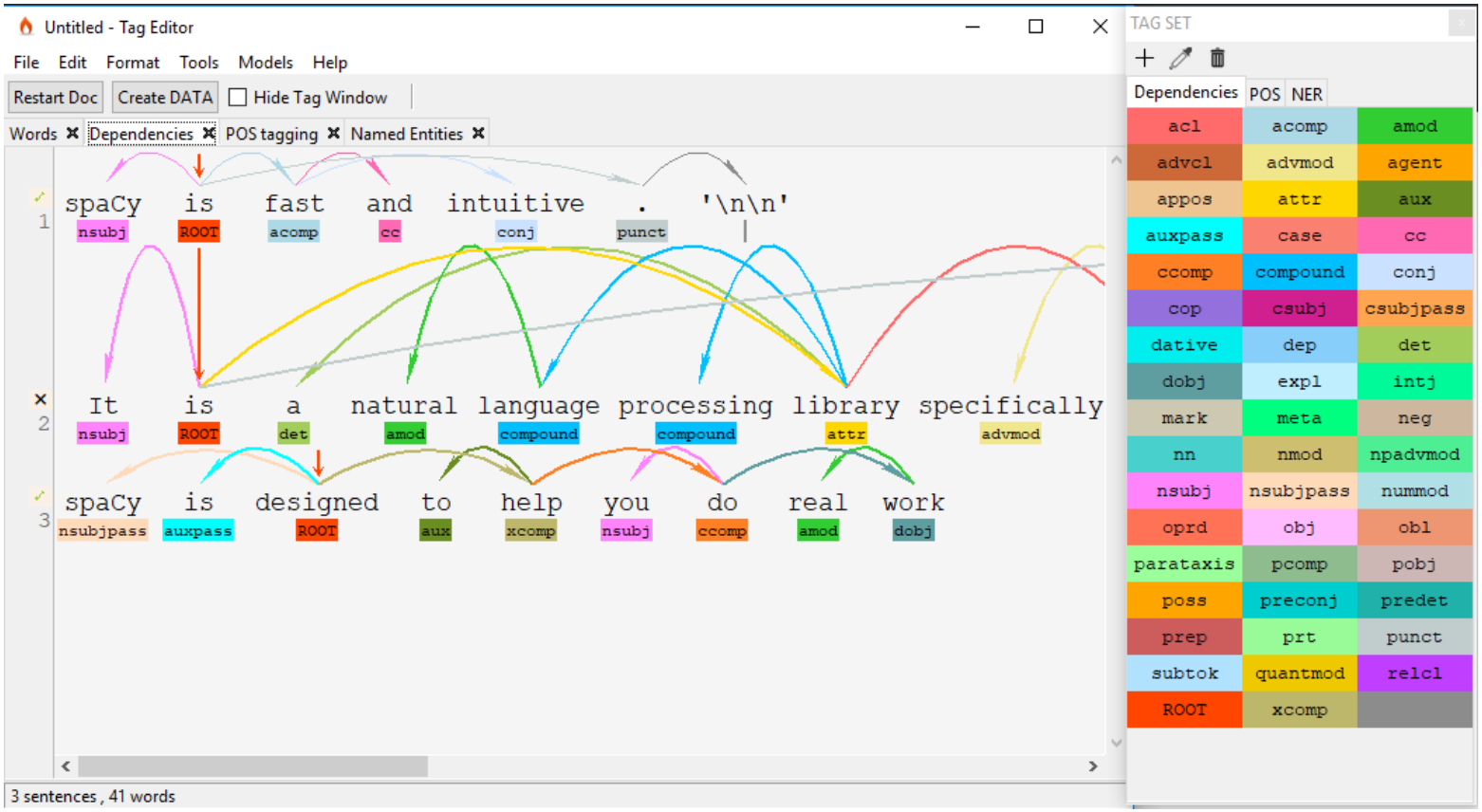


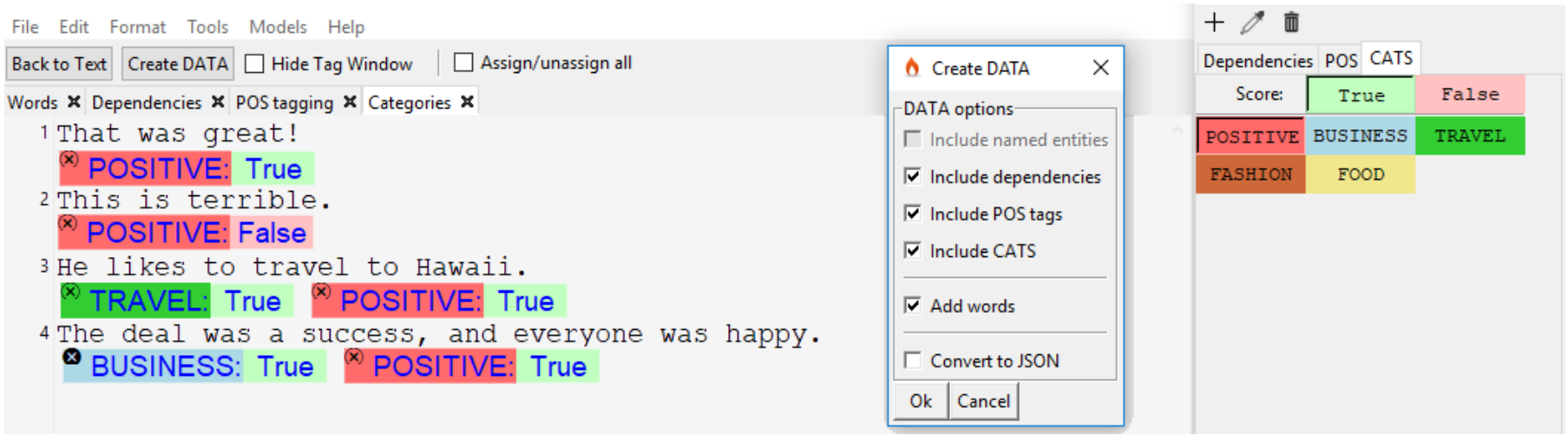
Analyse de Corpus

- Construire, vérifier et appliquer des dictionnaires électroniques
- Appliquer les tables du lexique-grammaire
- Aligner des textes
- Traiter l'ambiguité 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


GATE
general architecture
for text engineering
Products
Services
Support
Download
Documentation
Contact
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 rig
GATE is an open source software toolkit capable of solving almost any text process
It has a mature and extensive community of developers, users, educators, students
It is used by corporations, SMEs, research labs and Universities worldwide
File Options Tools Help
ATE
Applications
Language Resource
test-ontology-b
voting-example-
Processing Resourc
Datastores
MimeType
gate.SourceURL
Messages
voting-example-...
test-ontology-b...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
OAT
RATC
RATI
Text
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
Filter:
New Inst.
Add to Selected Inst.
Instance
Label
Property
Value
Oxfordshire
[Oxfordshire]
partof
[Location]
Witney
[Witney]
partof
Oxfordshire
test-ontology-bdm-instar
Entity
Location
Country
Organization
Person
Leader
Key
Document Editor
Initialisation Parameters

https://github.com/d5555/TagEditor





prodigy

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

https://prodi.gy/

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 وبريطانيةوبولندية . لكن سرعان ما بات لزاماً على الهولنديين بنهاية الحرب الأوروبية أن يحاربوا مقاتلي الثورة الوطنيةالإندونيسية

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.jsonl

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

I live in New York CITY .

100%

_core_web_sm
,ORG

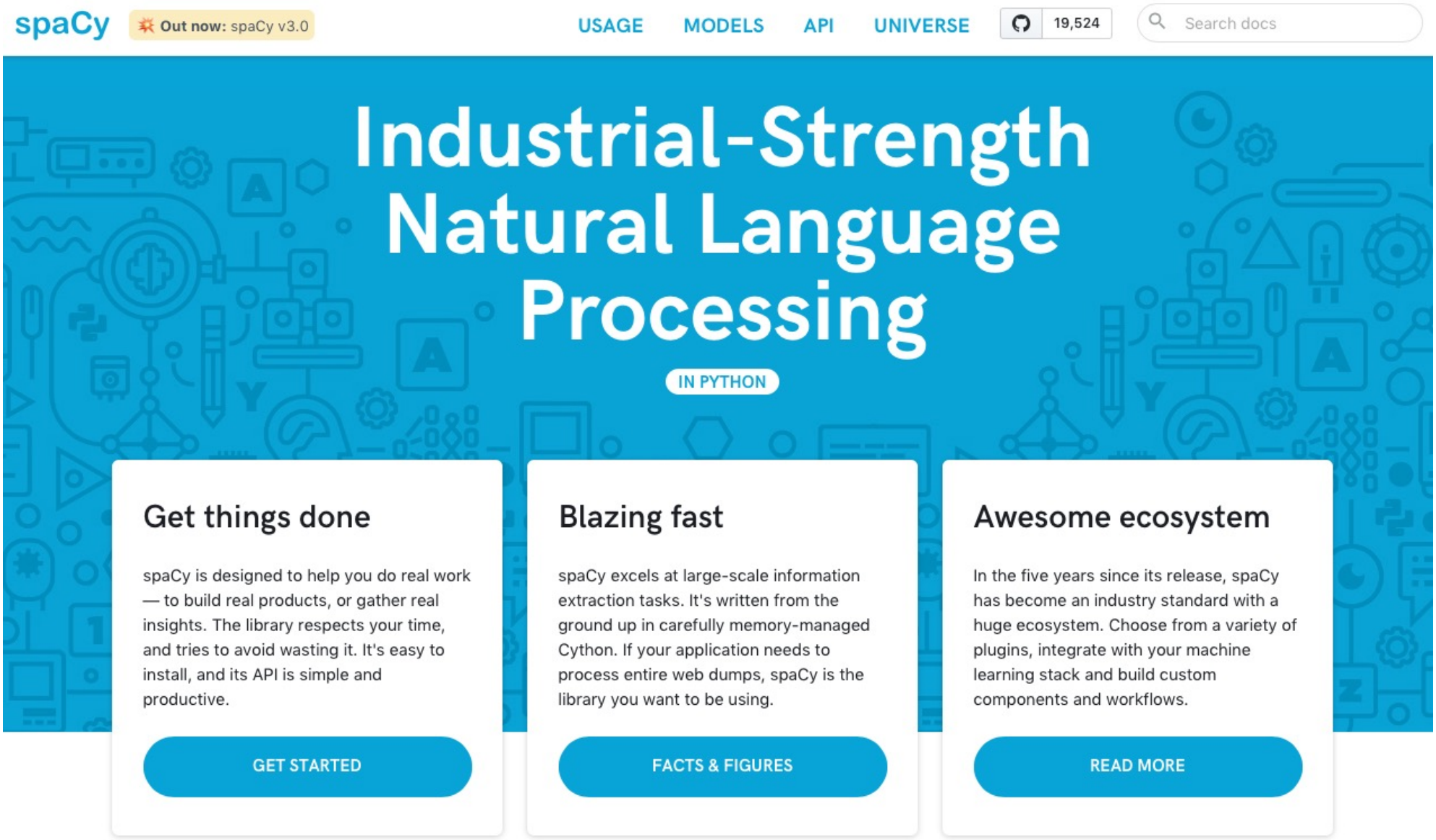
8080...
art annotating!

scratch binary

designer
as an
uality, life-

Spacy

https://spacy.io/



Spacy

https://spacy.io/

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

Fonctionnalités de Spacy

https://spacy.io/

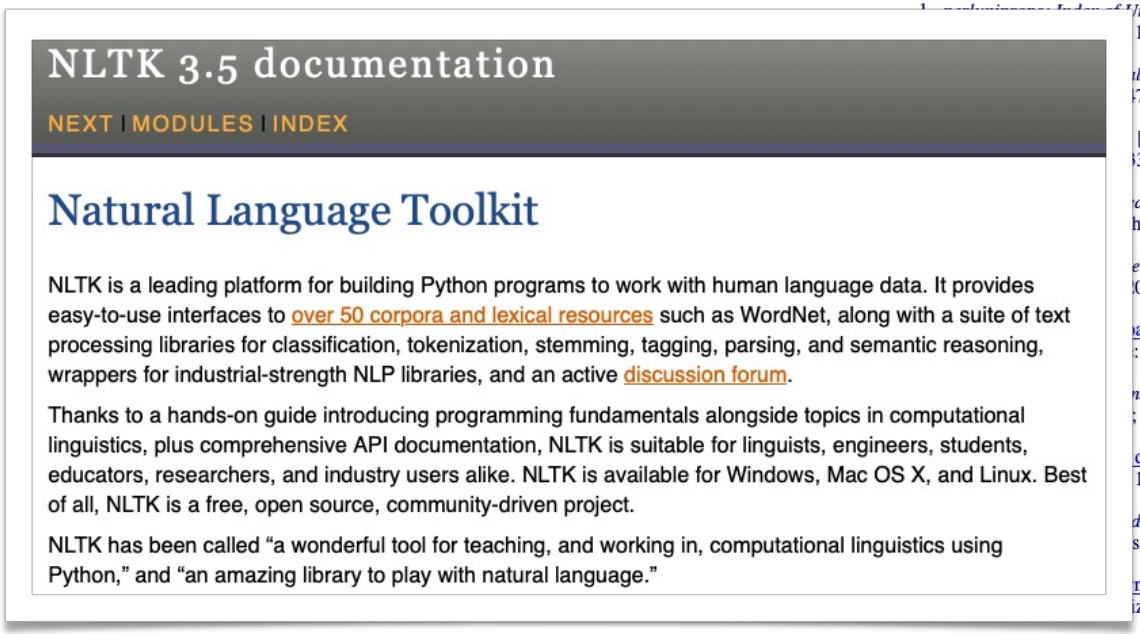
Features

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

http://www.nltk.org



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.”

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

Please consult the README file included with each corpus for further information.

Unicode Version 7.0.0 character properties in Perl [download | source]
100266; author: ; copyright: ; license: ;

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

[download | source]
3; author: Jan Strunk; copyright: ; license: ;

dor de Sufixos da Lingua Portuguesa) [download | source]
hor: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;

es [download | source]
00510; author: ; copyright: ; license: ;

ad | source]
: 6785405; author: ; copyright: ; license: ;

nker (Maximum entropy) [download | source]
; size: 13404747; author: ; copyright: ; license: ;

download | source]
10961490; author: ; copyright: ; license: ;

del [download | source]
size: 24516205; author: ; copyright: ; license: ;

nload | source]
ze: 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: ;

NLTK

http://www.nltk.org

Some simple things you can do with NLTK

Tokenize and tag some text:

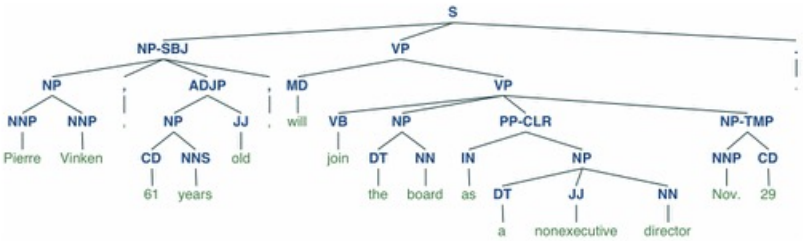
```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)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



NLTK

```
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:

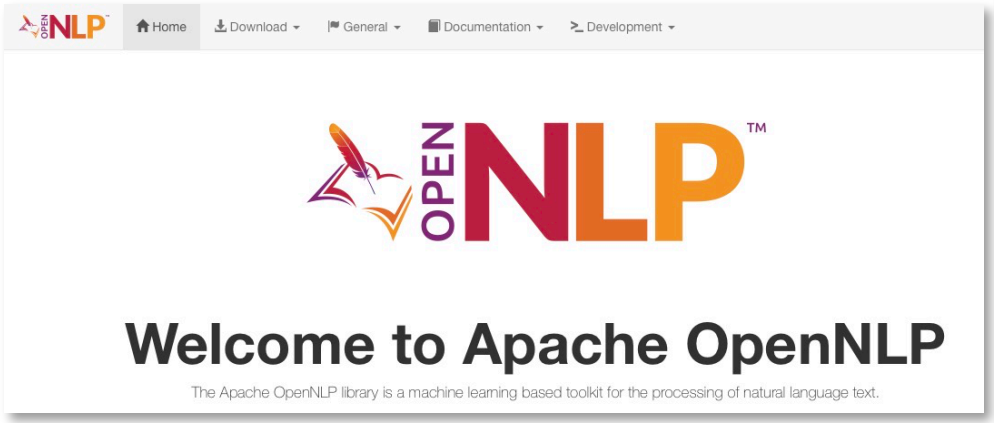
```
>>> training_set = sentim_analyzer.apply_features(training_docs)
>>> test_set = sentim_analyzer.apply_features(testing_docs)
```

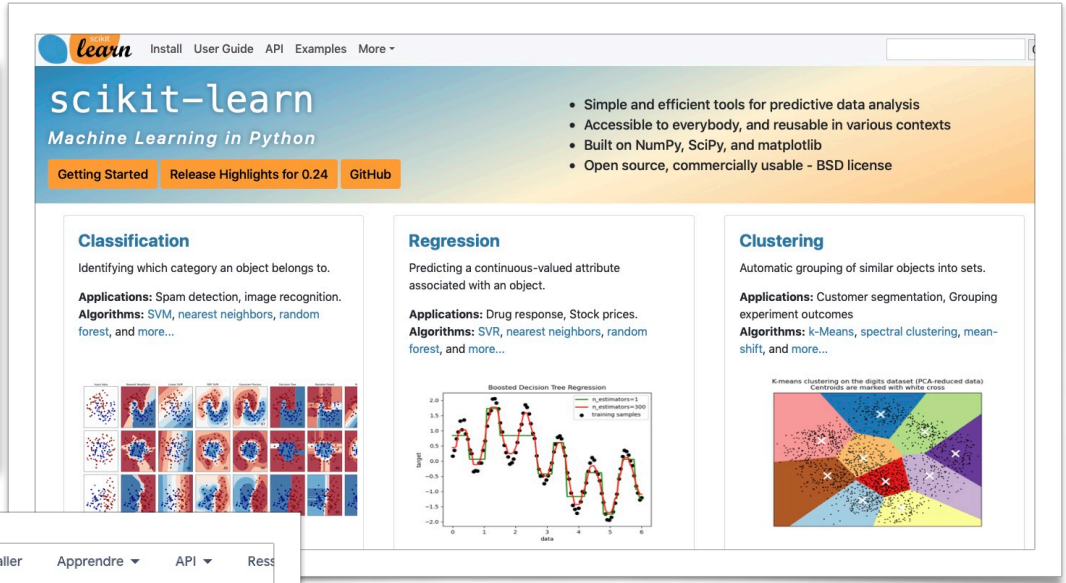
We can now train our classifier on the training set, and subsequently output the evaluation results:

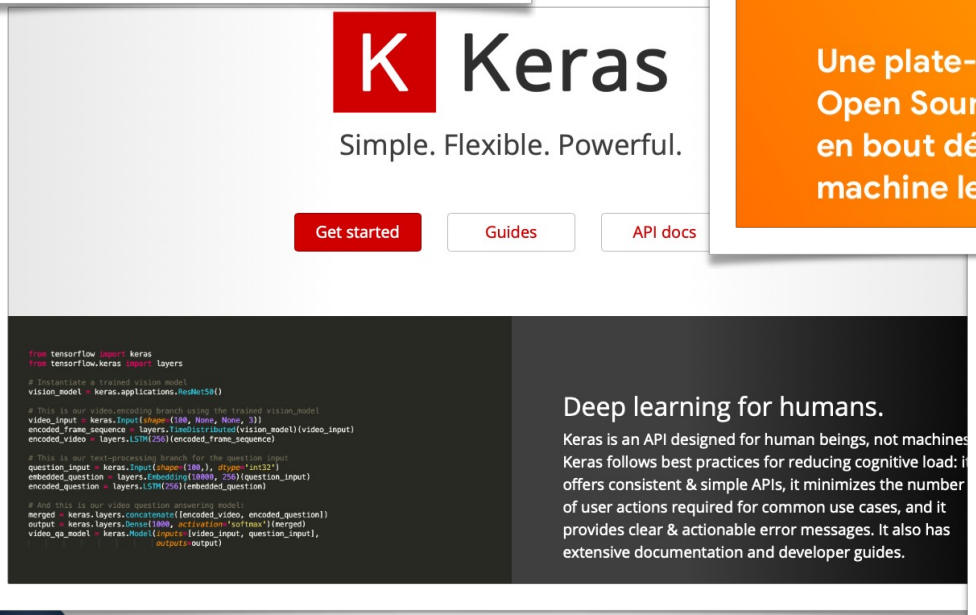
```
>>> trainer = NaiveBayesClassifier.train
>>> classifier = sentim_analyzer.train(trainer, training_set)
Training classifier
>>> for key,value in sorted(sentim_analyzer.evaluate(test_set).items()):
...     print('{0}: {1}'.format(key, value))
Evaluating NaiveBayesClassifier results...
Accuracy: 0.8
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
```

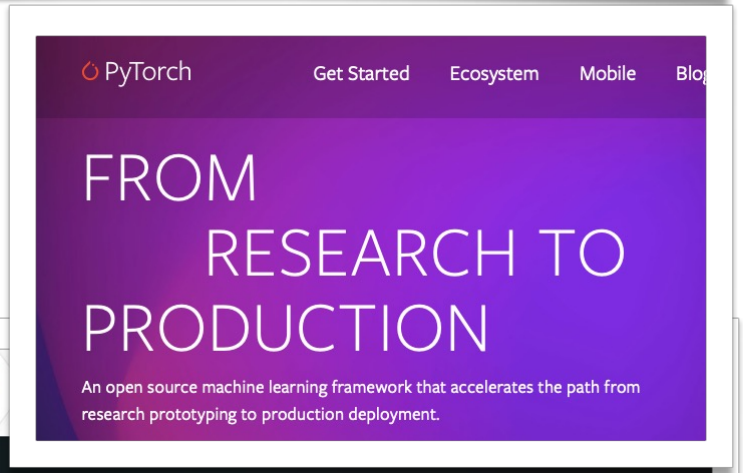
http://www.nltk.org

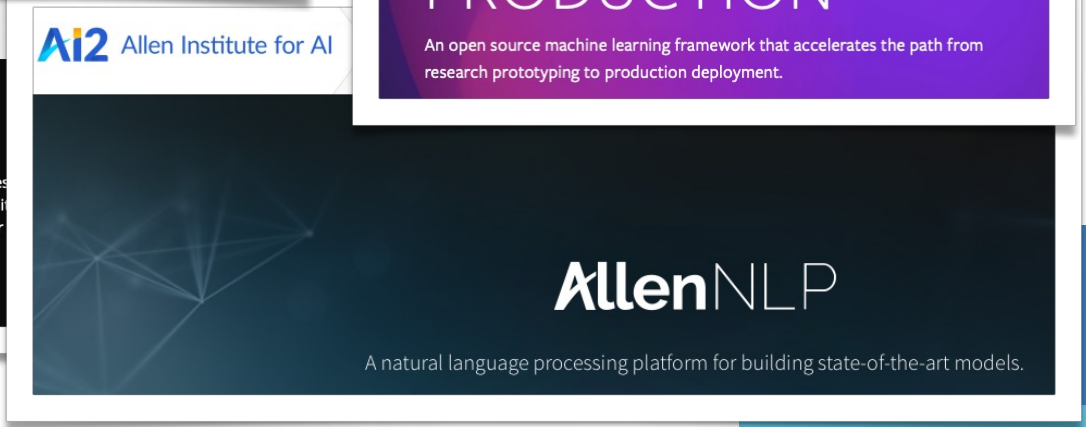
Des bibliothèques Python (ou Java, C++, Swift)











tm - Text Mining Package

http://tm.r-forge.r-project.org/

tm (shorthand for *Text Mining Infrastructure in R*) provides a framework for text mining applications within R.

The tm package offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R. The package has integrated database back-end support to minimize memory demands. An advanced meta data management is implemented for collections of text documents to alleviate the usage of large and with meta data enriched document sets.

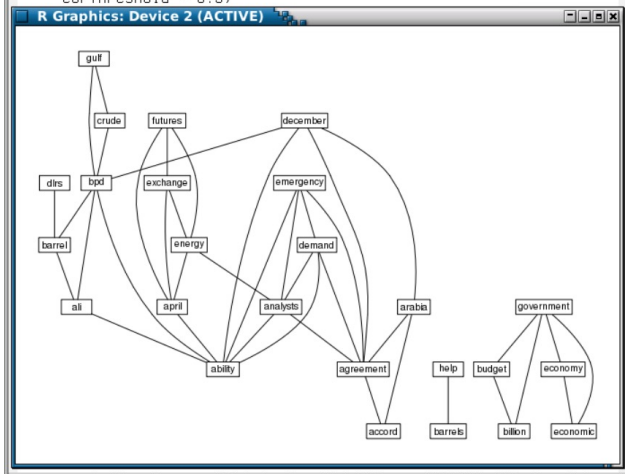
The package provides native support for reading in several classic file formats (e.g. plain text, PDFs, or XML files). There is also a plug-in mechanism to handle additional file formats.

The data structures and algorithms can be extended to fit custom demands, since the package is designed in a modular way to enable easy integration of new file formats, readers, transformations and filter operations.

tm provides easy access to preprocessing and manipulation mechanisms such as whitespace removal, stemming, or stopword deletion. Further a generic filter architecture is available in order to filter documents for certain criteria, or perform full text search. The package supports the export from document collections to term-document matrices.

tm is freely available under the GNU General Public License (GPL).

```
emacs@scorpia
File Edit Options Buffers Tools iESS Complete In/Out Signals Help
library("tm")

data("crude")

tdm <- TermDocumentMatrix(crude,
          control = list(removePunctuation = TRUE,
                         removeNumbers = TRUE,
                         stopwords = TRUE))

plot(tdm,
     terms = findFreqTerms(tdm, lowfreq = 6)[1:25],
     corThreshold = 0.5)
```




```
--:---  plotTermDocumentMatrix.R   All L13    (ESS[S] [R] Rox Abbrev)
```

openNLP: Apache OpenNLP Tools Interface

An interface to the Apache OpenNLP tools (version 1.5.3). The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java. common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. See <https://opennlp.a information.

Version:	0.2-7
Imports:	NLP (≥ 0.1-6.3), openNLPdata (≥ 1.5.3-1), rJava (≥ 0.6-3)
Suggests:	openNLPmodels.en
Published:	2019-10-26
Author:	Kurt Hornik [aut, cre]
Maintainer:	Kurt Hornik <Kurt.Hornik at R-project.org>
License:	GPL-3
NeedsCompilation:	no
SystemRequirements:	Java (>= 5.0)
In views:	NaturalLanguageProcessing
CRAN checks:	openNLP results

RSentiment: Analyse Sentiment of English Sentences

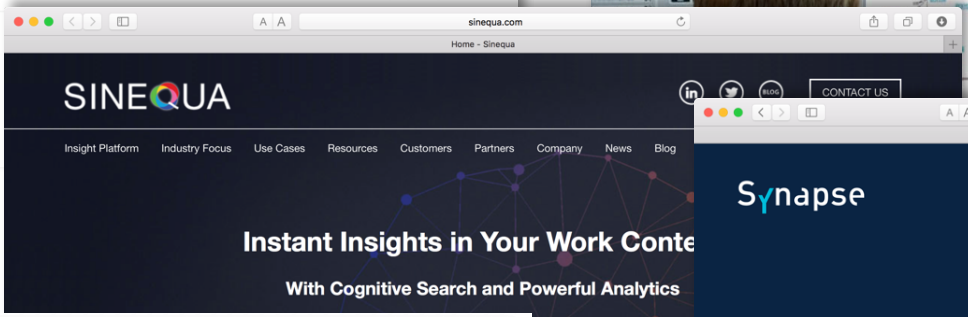
Analyses sentiment of a sentence in English and assigns score to it. It can classify sentences to the following categories of sentiments:- Positive, Negative, very Positive, very negative, Neutral. sentences, it counts the number of sentences in each category of sentiment.In calculating the score, negation and various degrees of adjectives are taken into consideration. It deals only with En

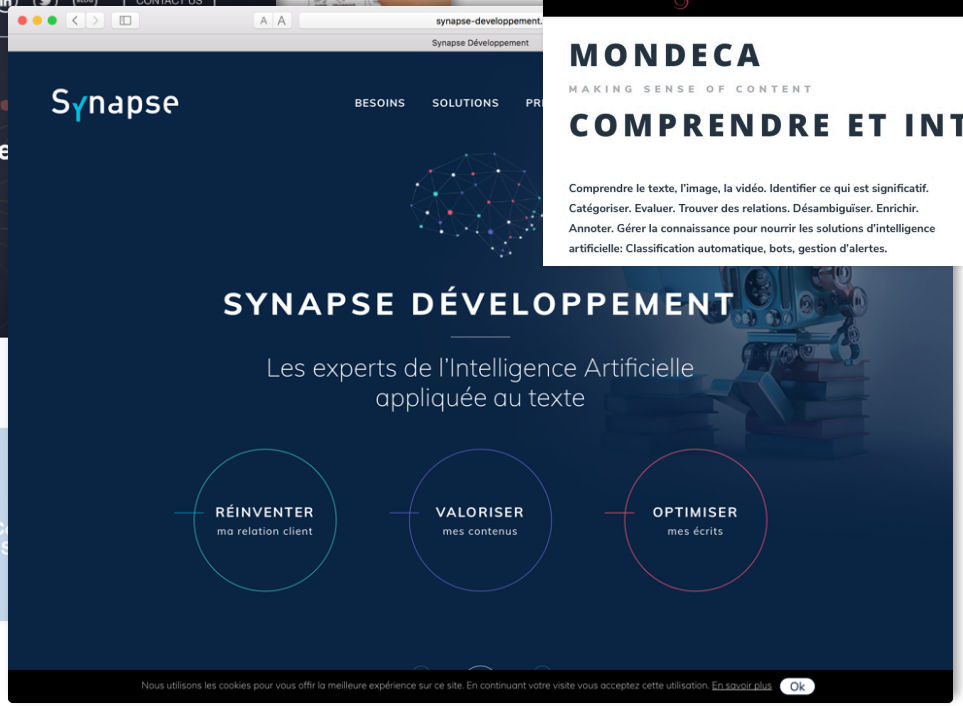
Version:	2.2.2
Depends:	R (≥ 2.1.0)
Imports:	plyr, stringr, openNLP, NLP
Suggests:	knitr, rmarkdown, testthat

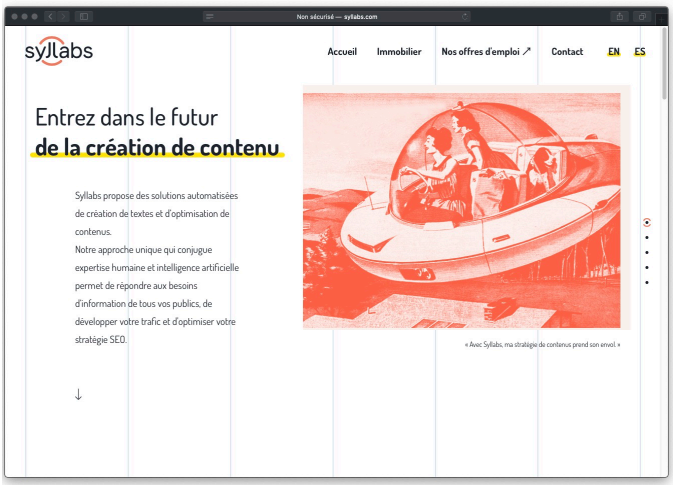
Des plateformes logicielles orientée « fouille de textes »

04.06.19

De nombreux acteurs industriels







Devon Think

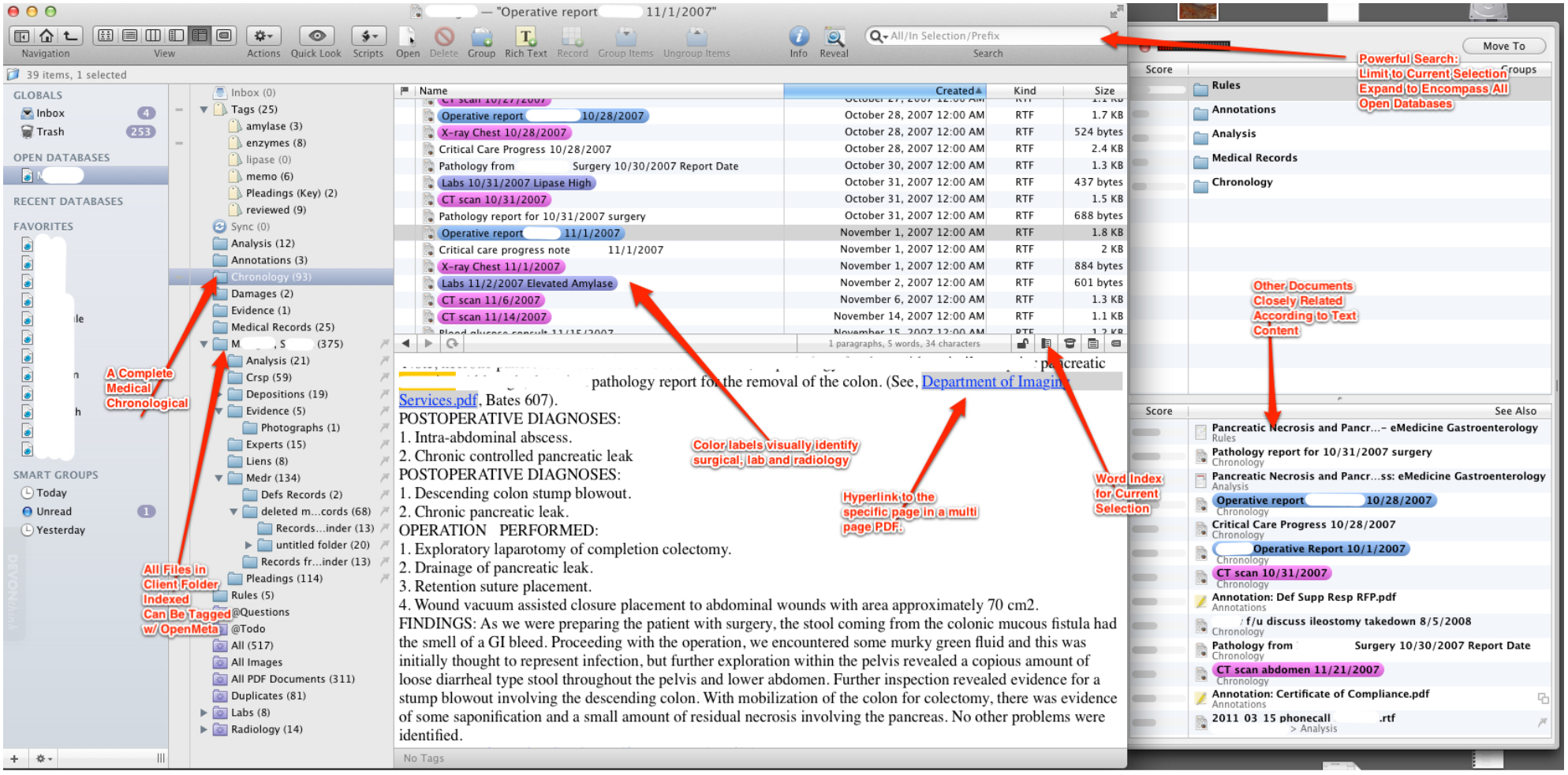
The DEVONtechnology consists of a small and efficient kernel and several abstraction layers. Its core functions include:

- 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:

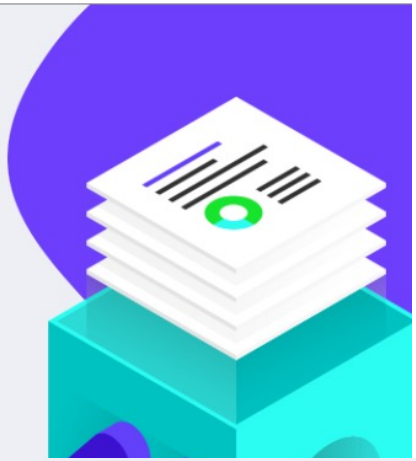
- Knowledge bases
- Expert systems
- Search engines
- Table-of-content generators
- Instant data-mining
- Intelligent agents
- Thesauri and context-sensitive help

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



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

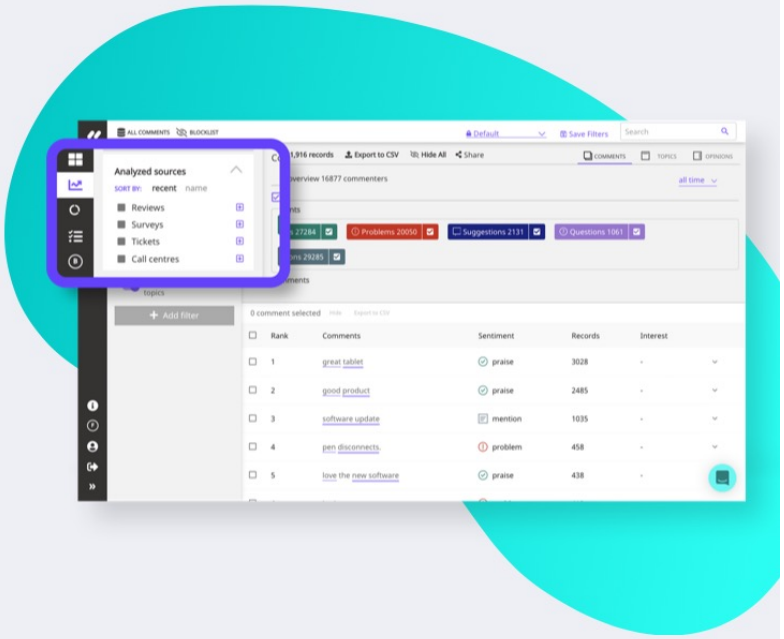
Instantly access the Praises, Problems, Suggestions, Questions, and Mentions in your feedback with accurate data categorization.

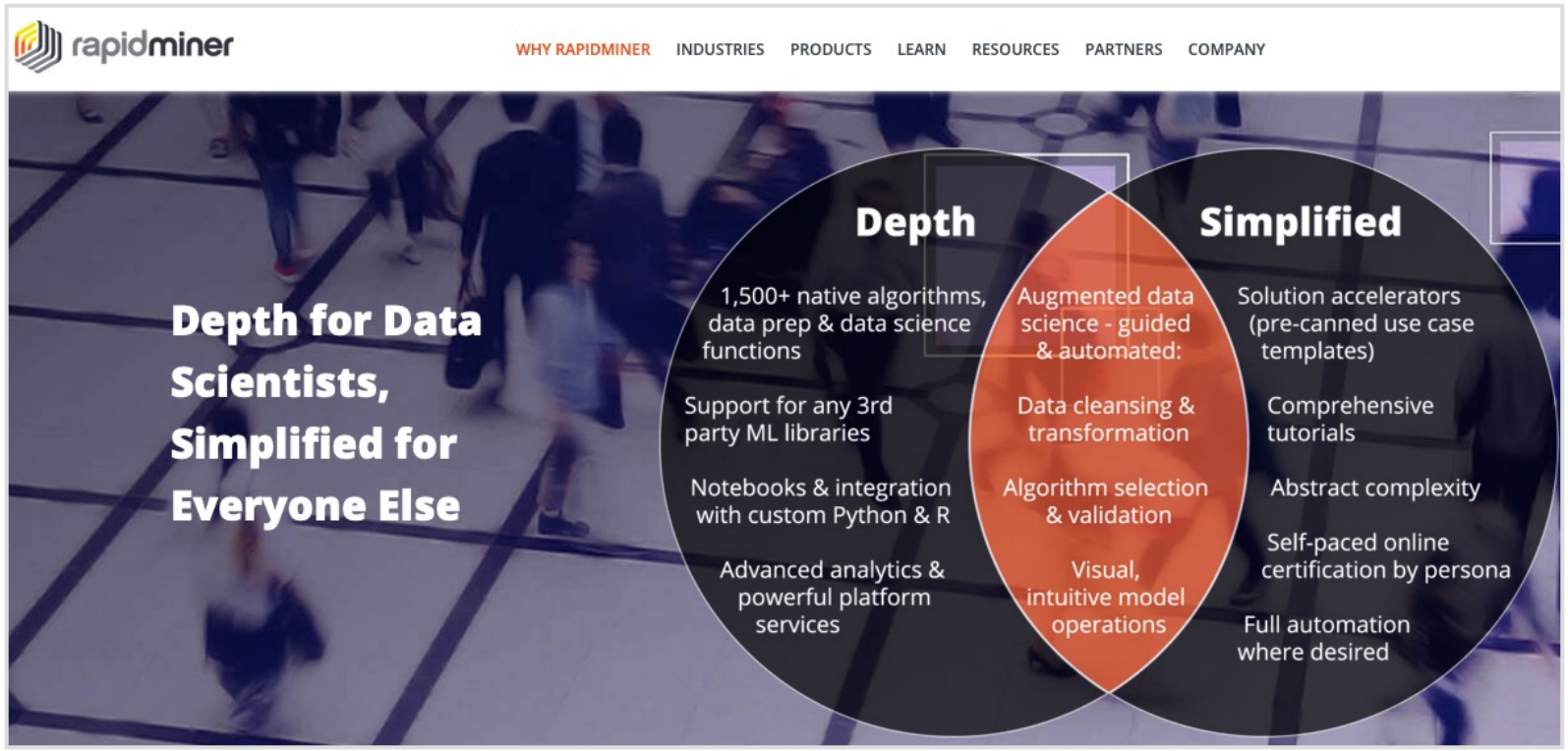
Multilingual processing

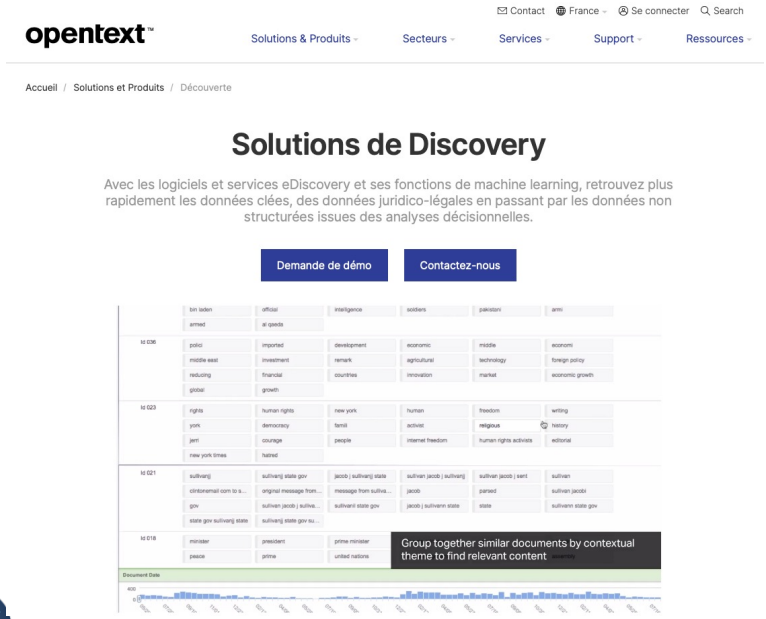
Analyze feedback with native proficiency in English and French, and top-of-the-line AI that translates 50+ other languages to English.

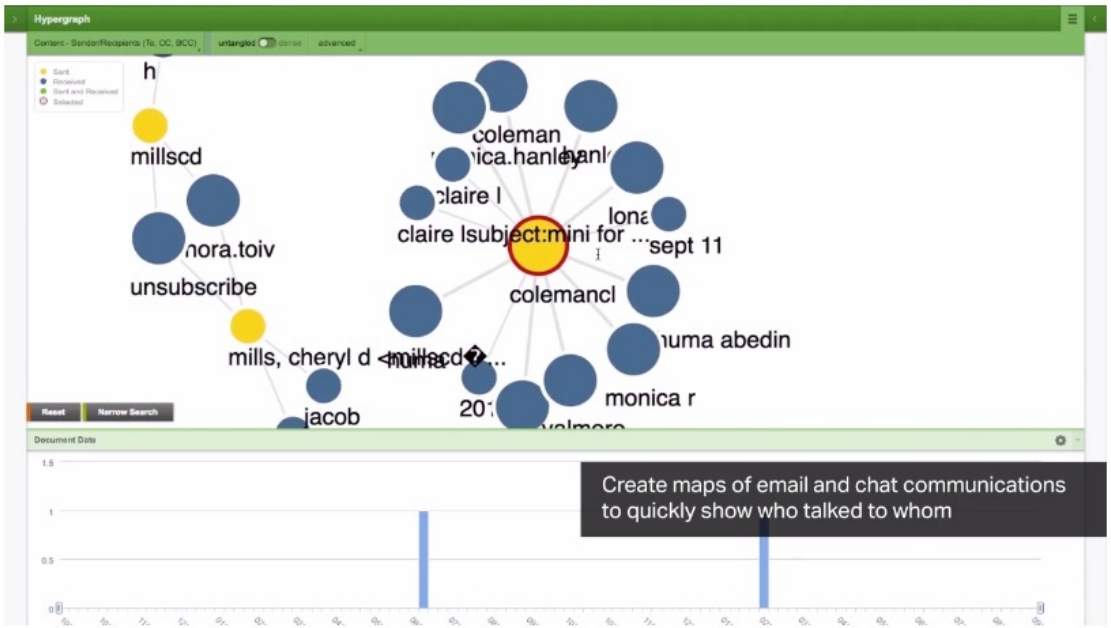
Advanced data filtering

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™
Contact
France
Se connecter
Search
Solutions & Produits
Secteurs
Services
Support
Ressources
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 clées, des données juridico-légales en passant par les données non structurées issues des analyses décisionnelles.
Demande de démo
Contactez-nous
Group together similar documents by contextual theme to find relevant content

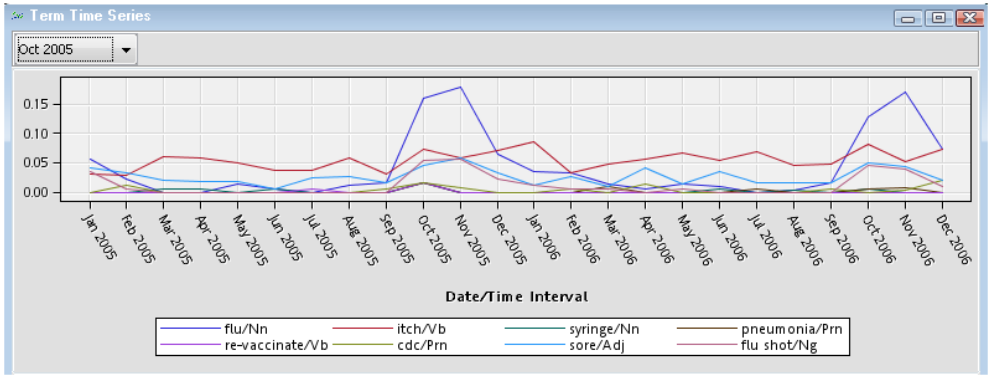

Hypergraph
millscd
nora.toiv
unsubscribe
mills, cheryl d
jacob
coleman
claire l
claire lsubject:mini for
sept 11
colemancl
monica r
Reset
Narrow Search
Document Date
Create maps of email and chat communications to quickly show who talked to whom

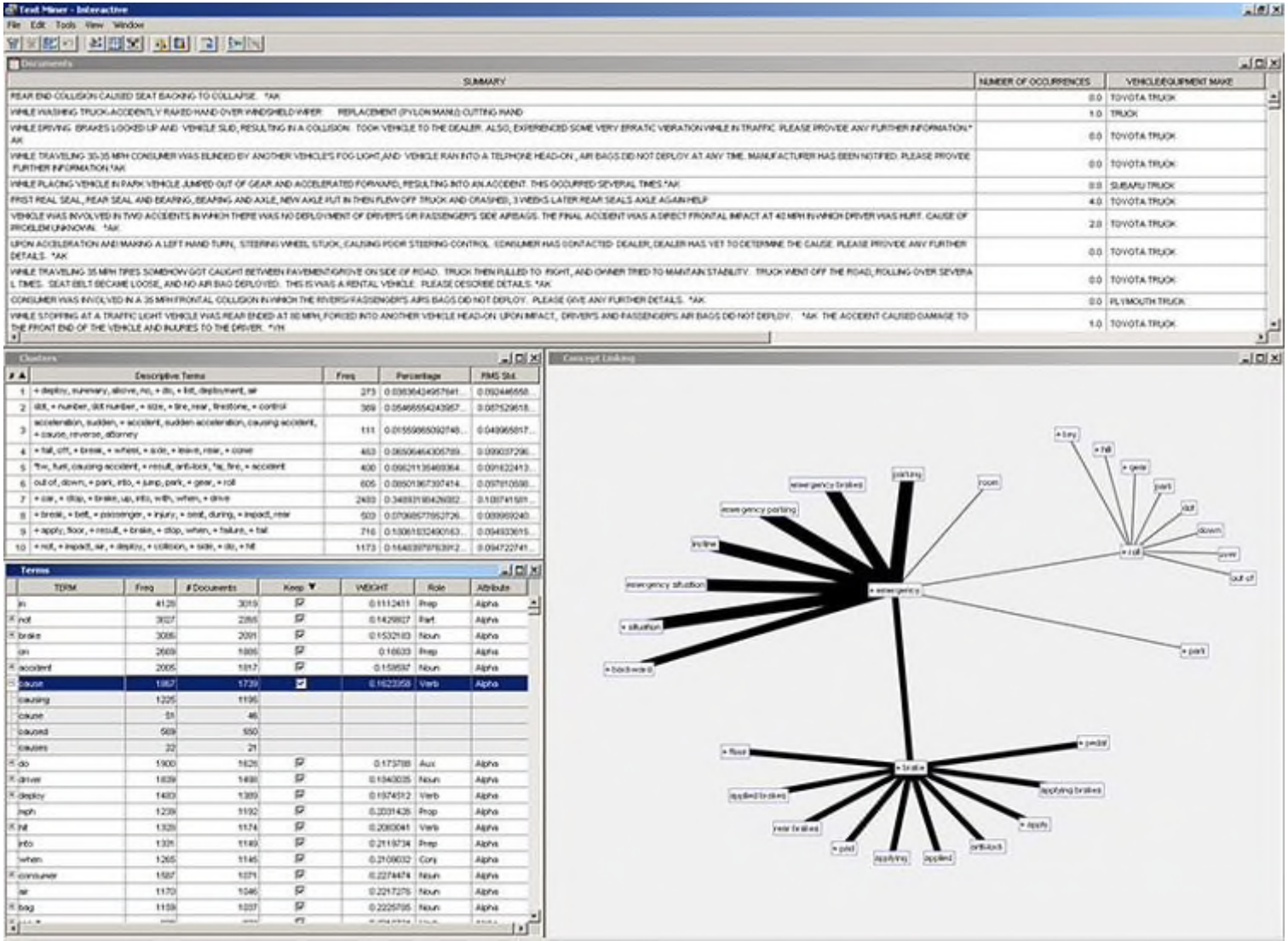
https://www.sas.com/fr_fr/software/text-miner.html

MENU

SAS® Text Miner

Mettez en évidence les informations dissimulées dans les données non structurées





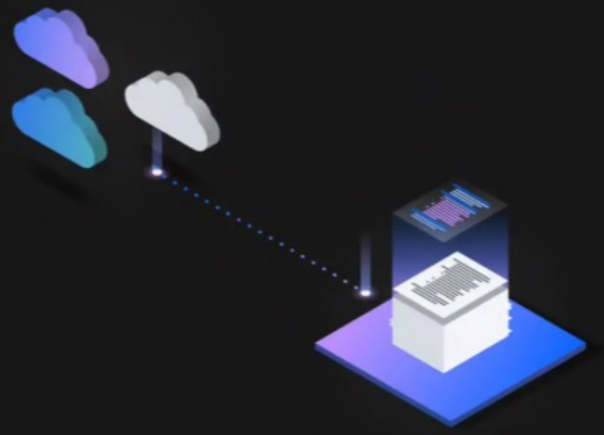
cnrs

https://www.ibm.com/cloud/watson-natural-language-understanding

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.

Overview

Powerful Insight Extraction

Get underneath the topics mentioned in your data by using text analysis to extract keywords, concepts, categories and more.

Learn more

Extensive Language Support

Analyze your unstructured data in more than thirteen languages.

Learn more

High-Accuracy Extraction

Out-of-the-box machine learning models for text mining provide a high degree of accuracy across your content.

Learn more

Why Watson NLU?

Deploy Anywhere

Deploy Watson Natural Language Understanding behind your firewall or on any cloud.

Learn more

Domain Customization

Train Watson to understand the language of your business and extract customized insights with Watson Knowledge Studio.

Learn more

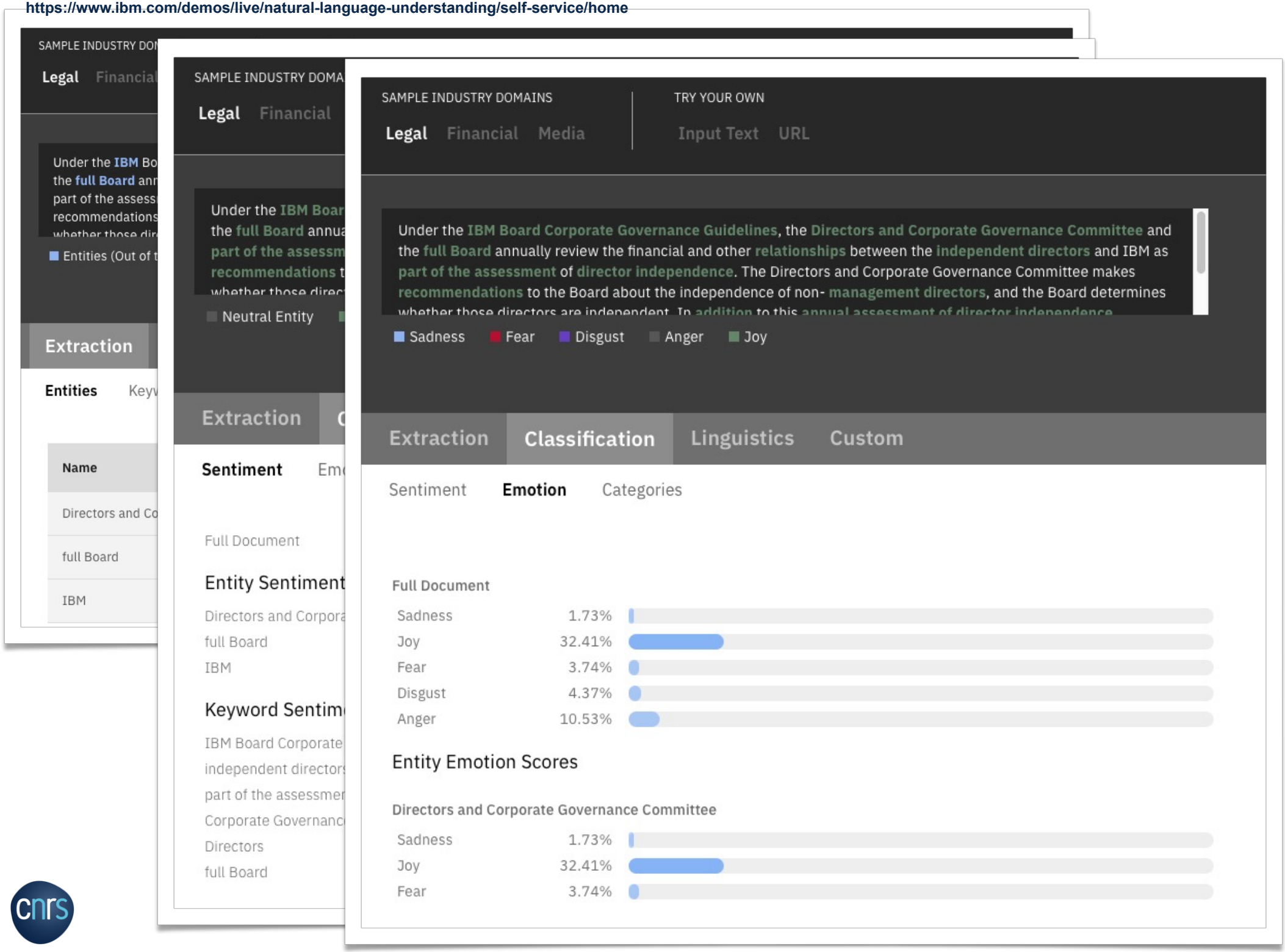
Data Control

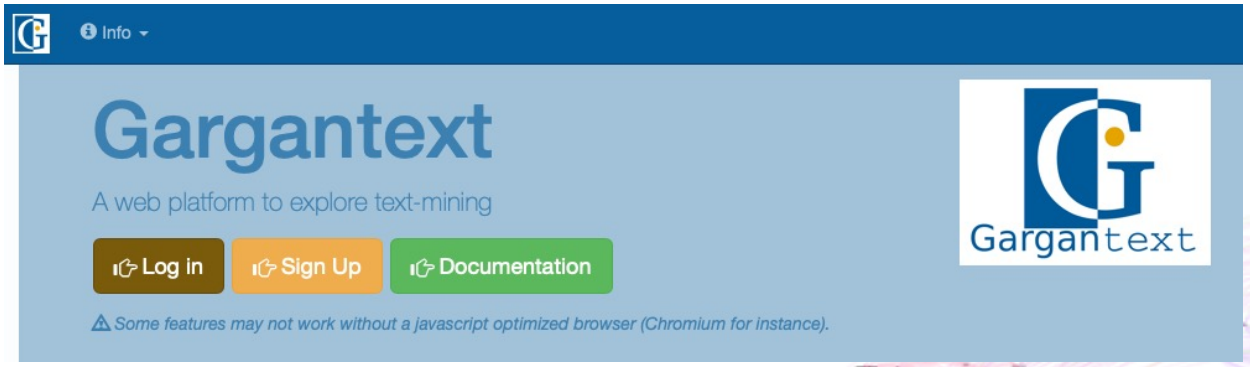
Maintain ownership of your data with the assurance that your data is safe and secure. IBM will not collect or store your data.

Learn more

Let's talk

https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

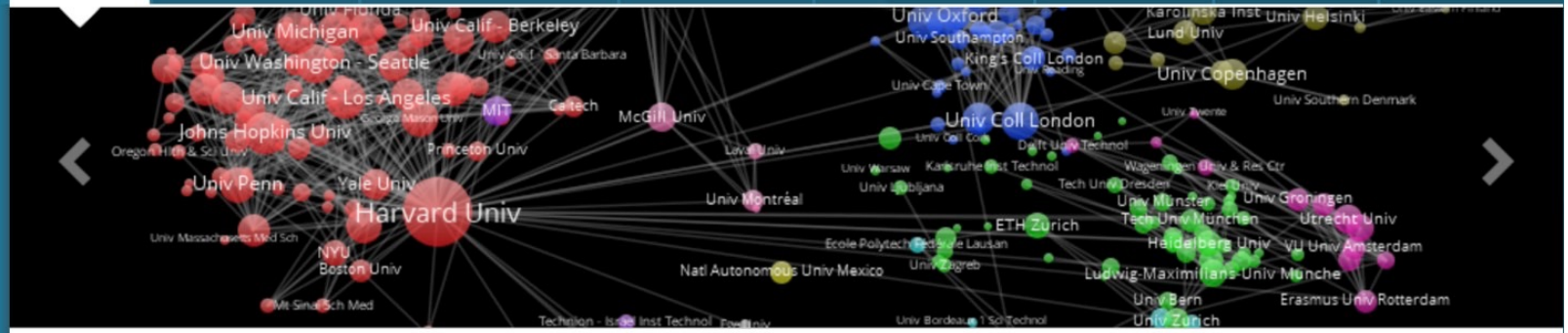




a collaborative web platform for the exploration of sets of unstructured documents. ***It combines tools from natural language processing, text-mining, complex networks analysis and interactive data visualization*** *to pave the way toward new kinds of interactions with your digital corpora. In few minutes,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

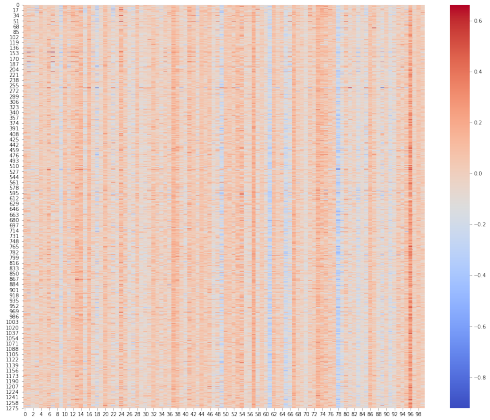
Click the button below to launch VOSviewer directly from this web page. This requires a system with Java support.

Launch VOSviewer ›

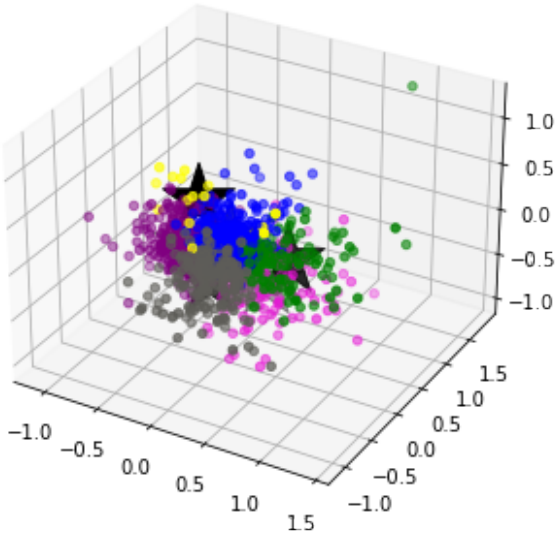

VOSviewer - operations_research_terms_map.txt - operations_research_terms_network.txt
Network Visualization
Density Visualization
Action
Items
Map
Filter:
schedul
16 items in 3 clusters:
Cluster 6 (14 items)
batch scheduling
feasible schedule
flowshop scheduling prol
job shop scheduling
job shop scheduling prob
line scheduling
machine scheduling prob
optimal schedule
project scheduling
project scheduling proble
scheduling
scheduling problem
single machine schedulin
single machine schedulin
Cluster 7 (1 item)
production schedule
Cluster 9 (1 item)
crew scheduling
Group items by cluster
data envelopment analysis
frontier consensus
multicriteria judgment
theoretical property
aggregation
norm
efficiency
output
organization
finding
layout
linear programming
research
scheduling problem
local search
graph
search
linear program
pilot
decade
information technolog
default
conclusion
asset
owner
machine
scheduling
node
convergence
market
offer
dynamic programming algorithm
travel time
note
planner
computational complexity
probability
consumer
game
supply side
system reliability
production process
price
promotion
random variable
pricing
item
service level
price compe
pricing policy
markov inventory
lot size
repair facility
inventory control
retail price
service station
replenishment
queue
inventory system
service time
inventory position
single server
permissible delay
markovian arrival process
retrial
Labels
Size:
Size variation:
Max. length:
30
Font:
Open Sans
Visualization
Item density
Cluster density
Kernel width:
Colors
White background
Density Colors...
Cluster Colors...

Visualisation ?

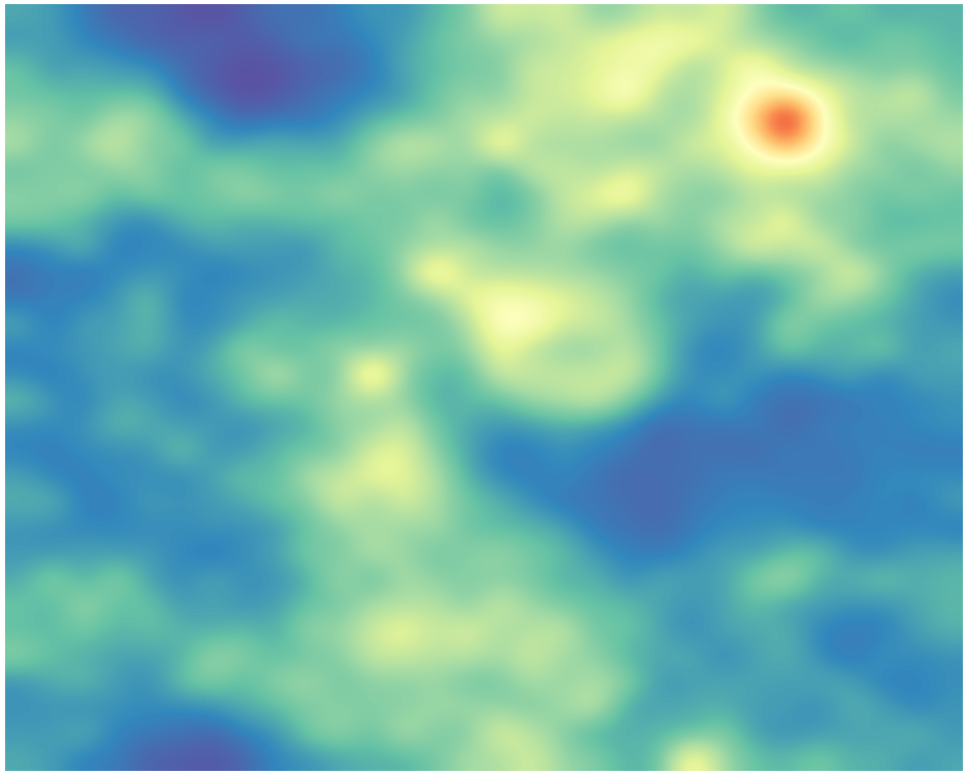
Représentation de documents par projection



Vectorisation dans des espaces continus (Doc2Vec) par plongements lexicaux



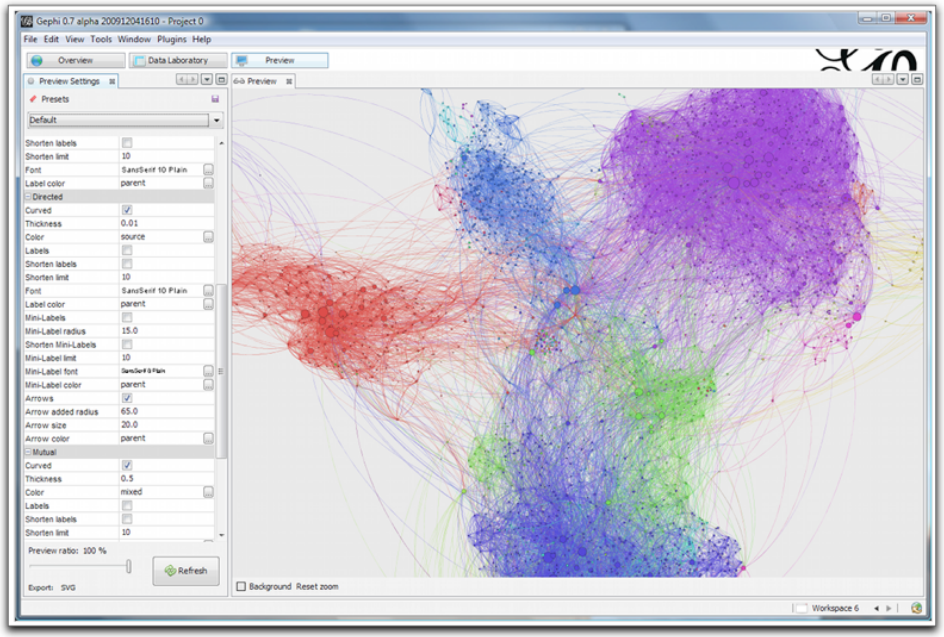
Analyse en Composantes Principales



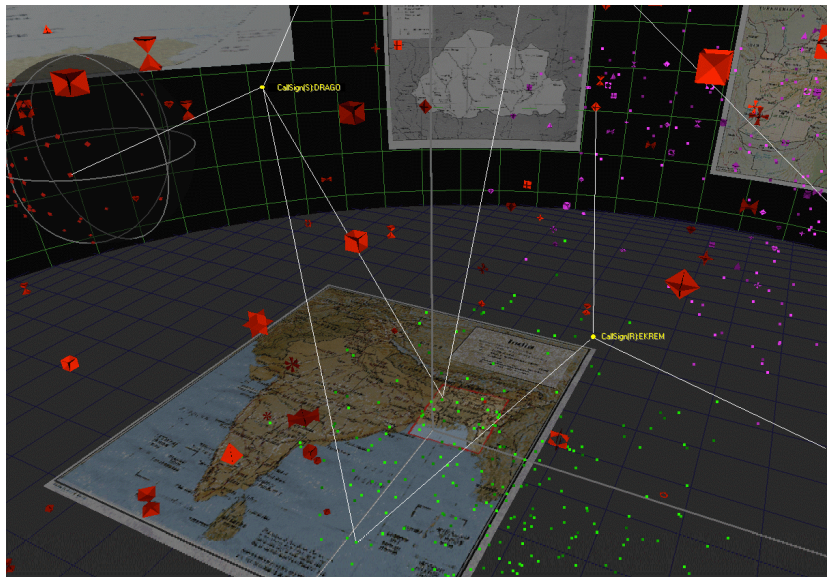
Cartes auto-organisées (SOM)

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

LuckySort



gephi CONSORTIUM



STARLIGHT Visual Information System™

newsmap

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 14, NO. 6, NOVEMBER/DECEMBER 2008

The Word Tree, an Interactive Visual Concordance

Martin Wattenberg and Fernanda B. Viégas

Abstract— We introduce the Word Tree, a new visualization and information-retrieval technique aimed at text documents. A word tree is a graphical version of the traditional "keyword-in-context" method, and enables rapid querying and exploration of bodies of text. In this paper we describe the design of the technique, along with some of the technical issues that arise in its implementation. In addition, we discuss the results of several months of public deployment of word trees on Many Eyes, which provides a window onto the ways in which users obtain value from the visualization.

Index Terms—Text visualization, document visualization, Many Eyes, case study, concordance, information retrieval, search.

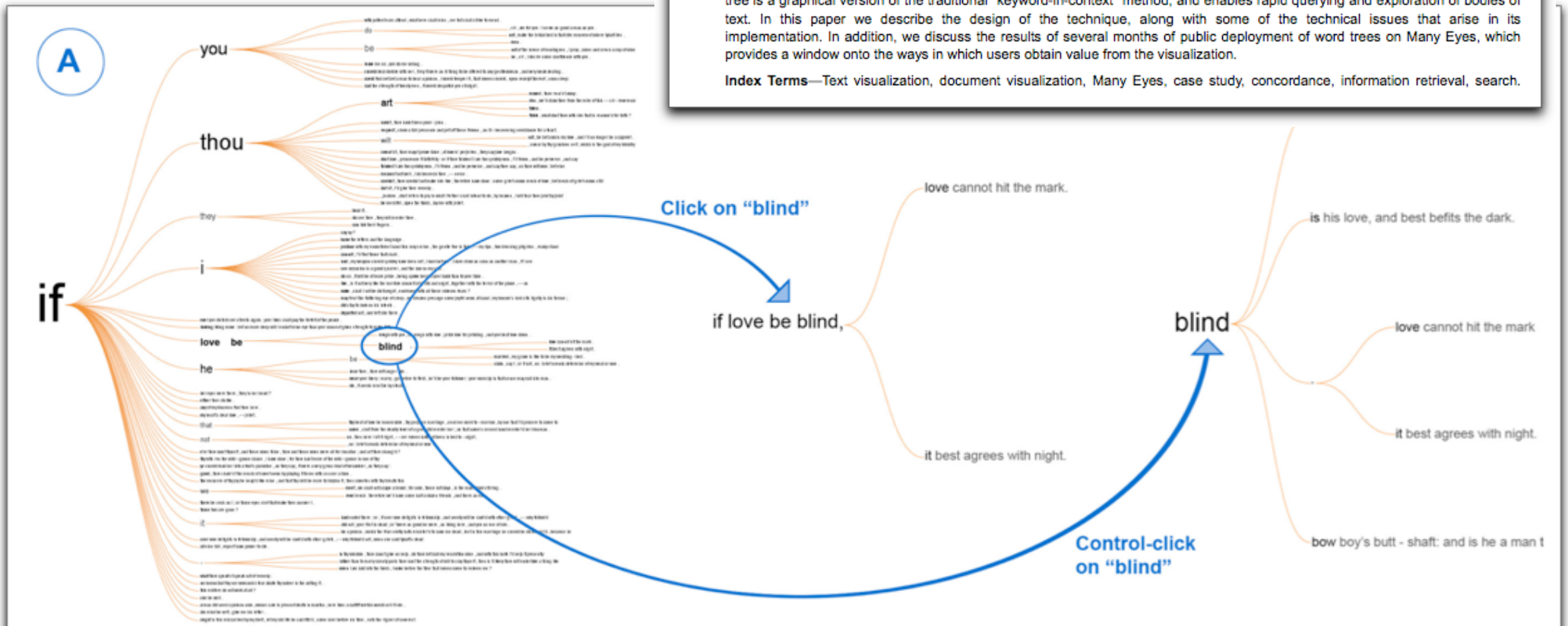
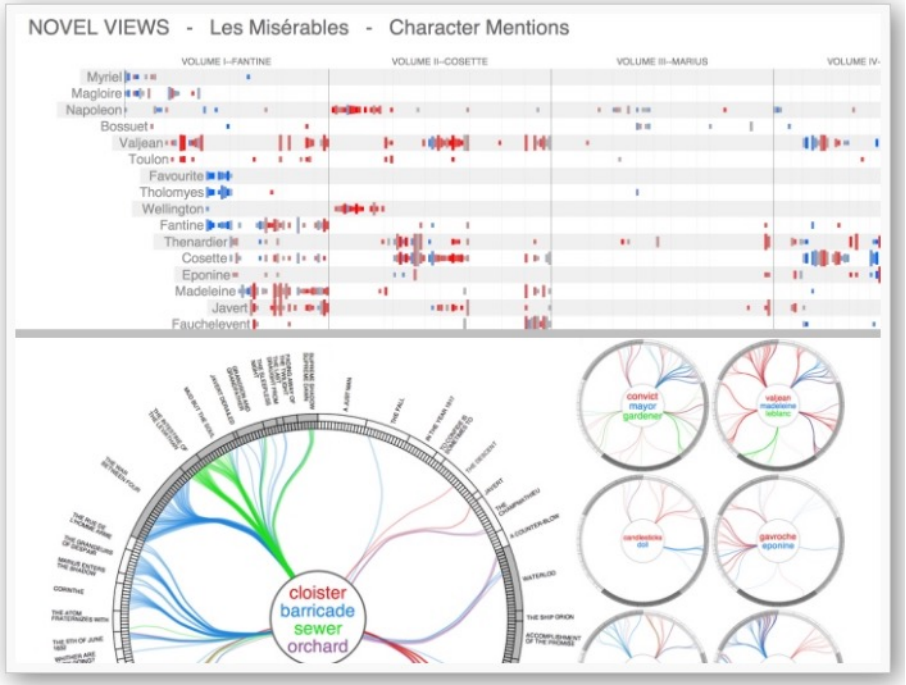
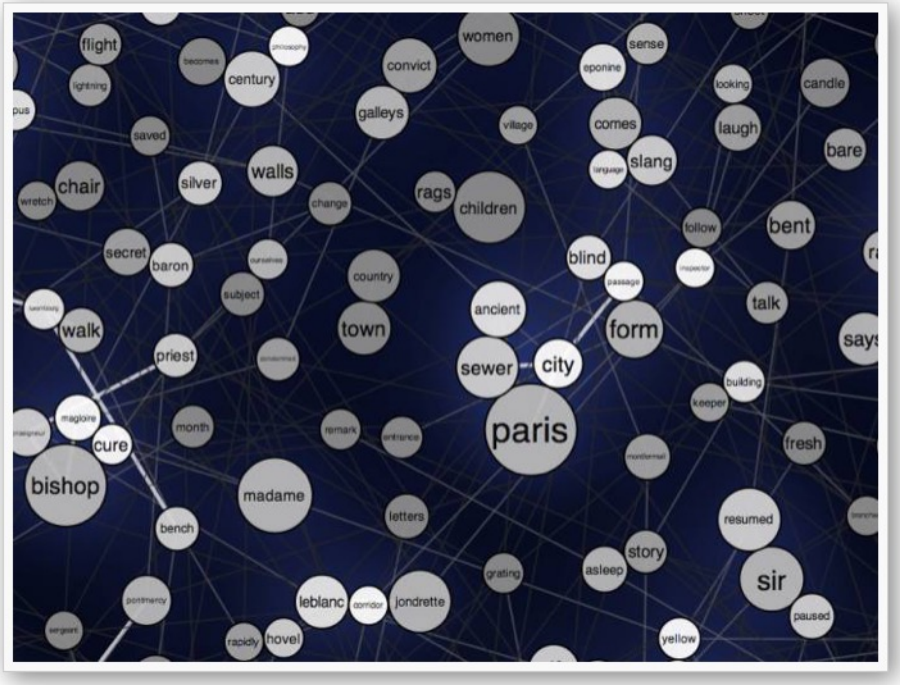


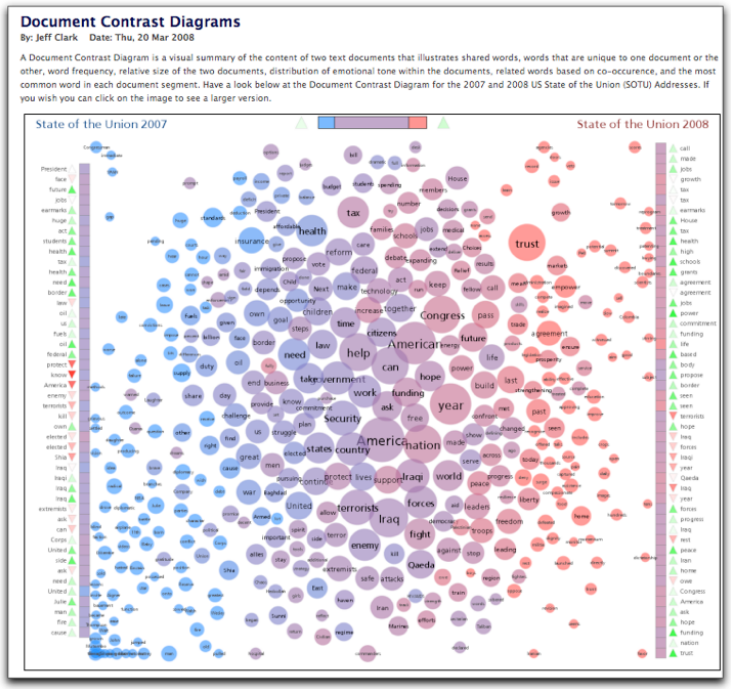
Fig 3. Sequence showing some of the interaction options in the word tree. In figure A, the user has typed the word “if” in *Romeo and Juliet*. In B, the user has clicked on “blind,” which appears in one of the branches under “if.” This causes the visualization to recenter to the longer phrase “if love be blind.” In C, the user Control-clicks on “blind,” which causes the visualization to recenter to blind by itself, revealing that there are additional phrases after this term.

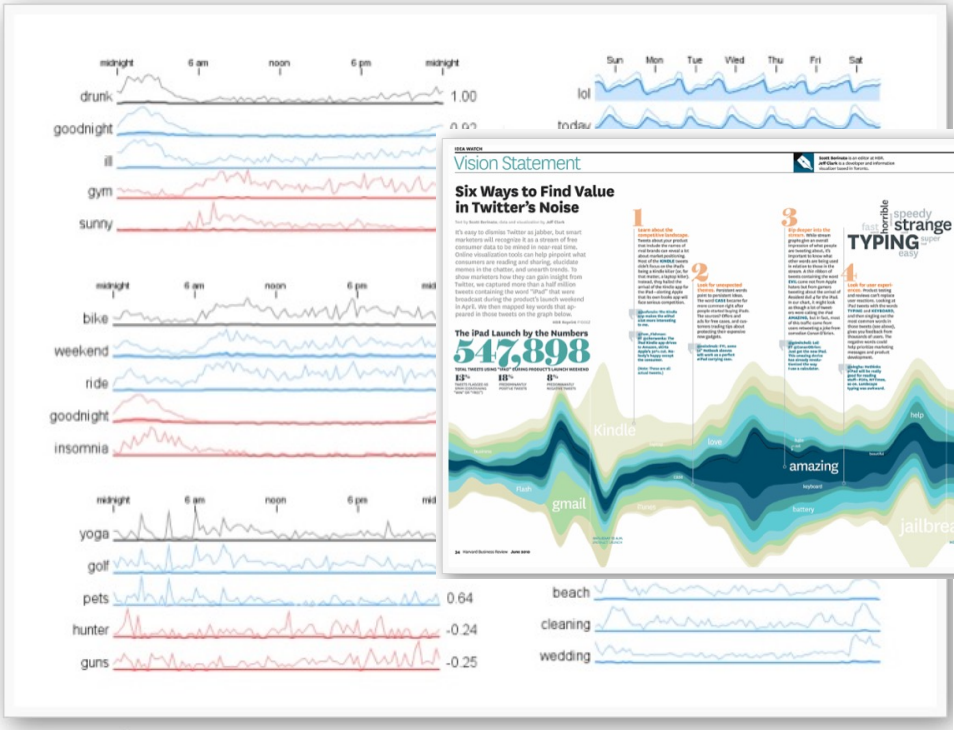
http://www-958.ibm.com/software/data/cognos/manyeyes/page/Word_Tree.html

Neoformix https://www.neoformix.com









Temporal Correlation for Words in Tweets

cnrs

GENSIM
topic modelling for humans

Home Documentation Support API About

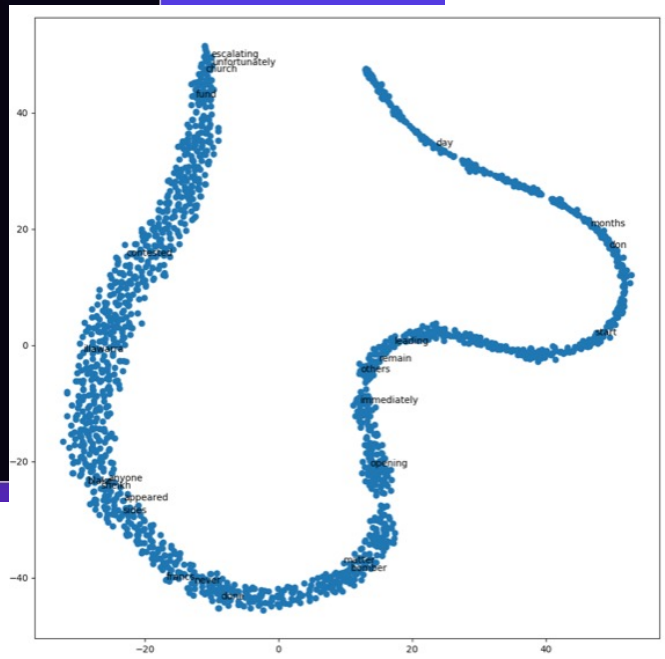
Fork on Github

Gensim is a FREE Python library

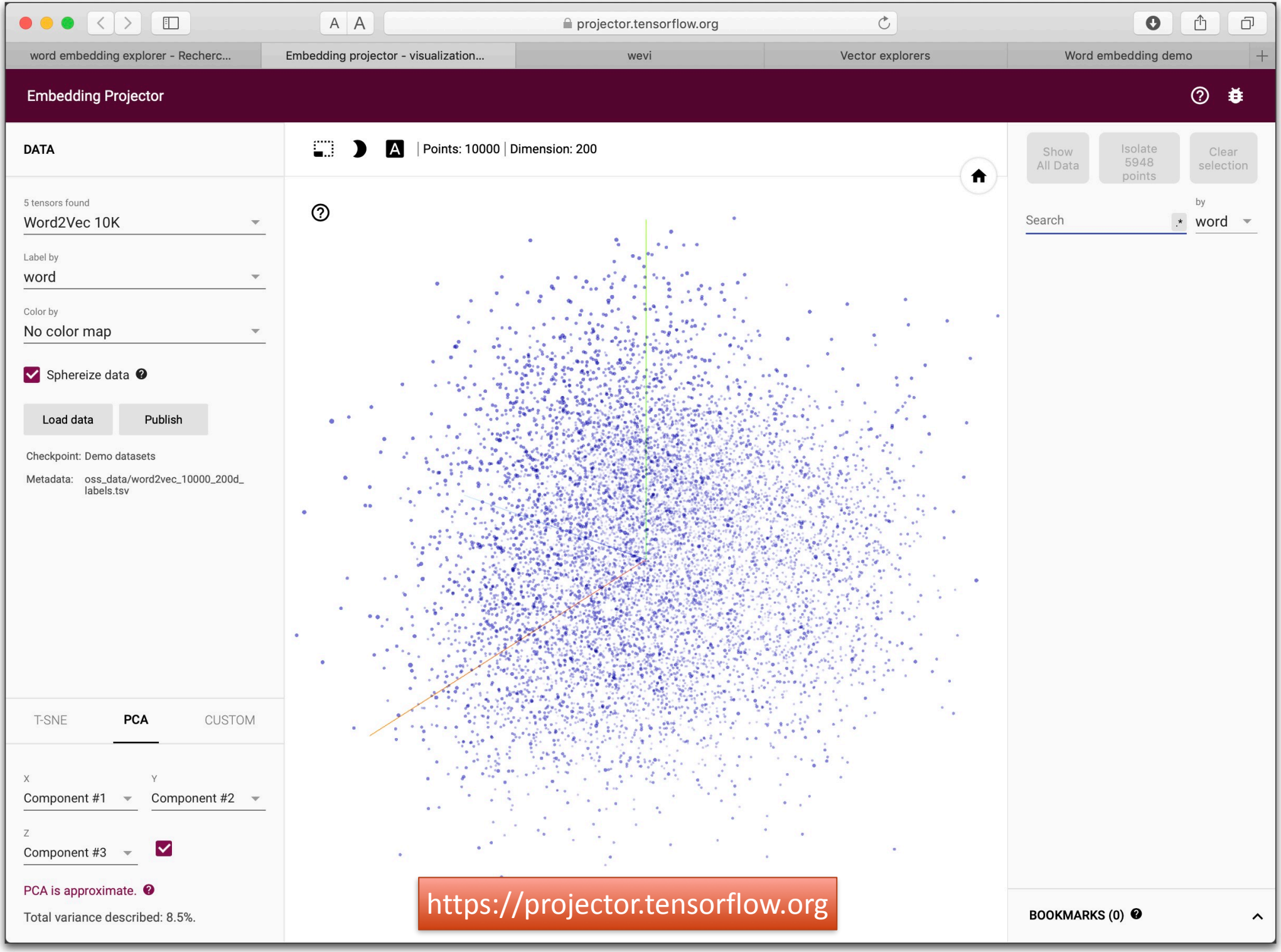
Topic modelling for humans

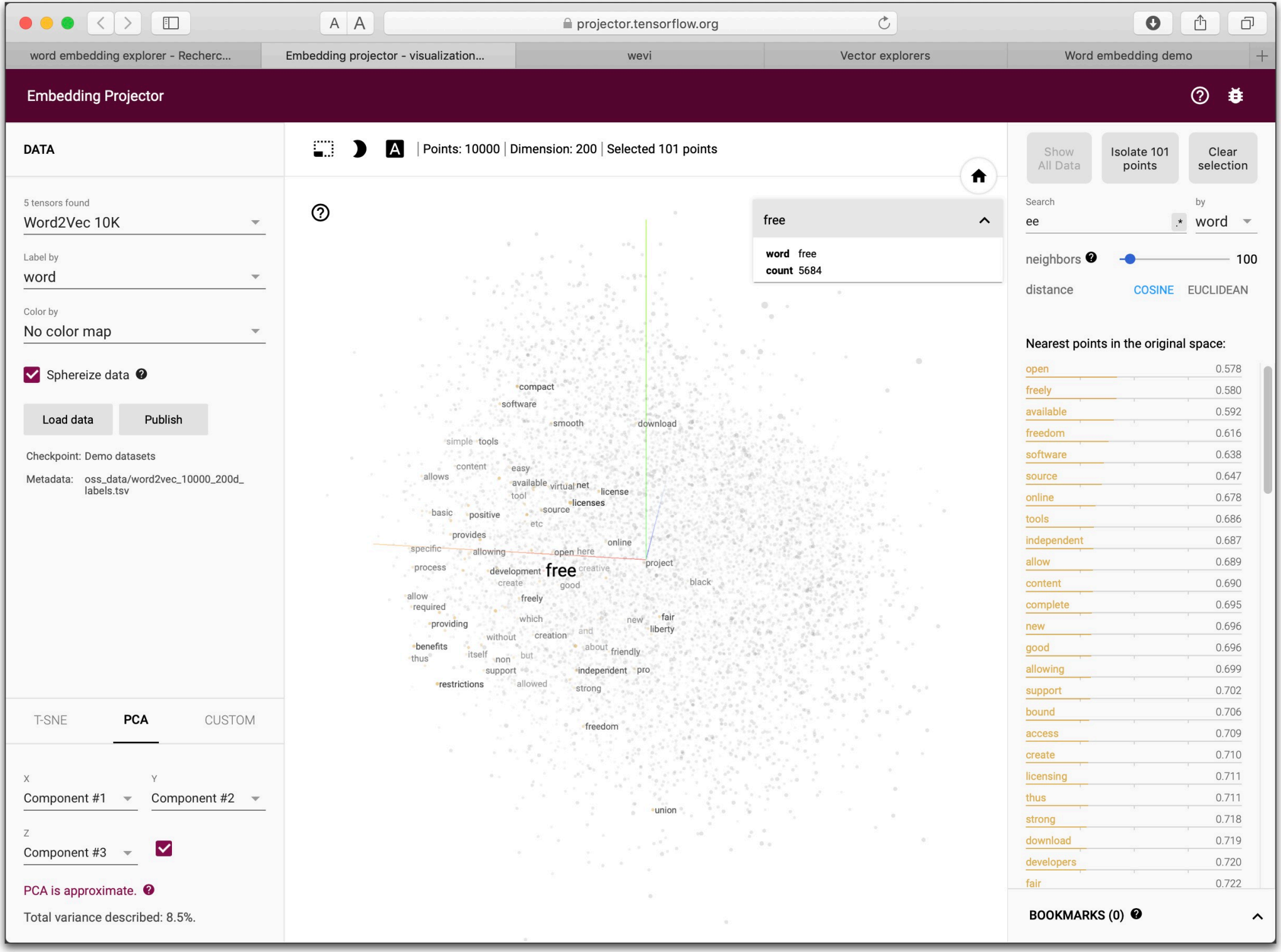
- ✔ Train large-scale semantic NLP models
- ✔ Represent text as semantic vectors
- ✔ Find semantically related documents

```
from gensim import corpora, models, similarities, downloader

# Stream a training corpus directly from S3.
corpus = corpora.MmCorpus("s3://path/to/corpus")

# Train Latent Semantic Indexing with 200D vectors.
lsi = models.LsiModel(corpus, num_topics=200)

# Convert another corpus to the LSI space and index it.
index = similarities.MatrixSimilarity(lsi[another_corpus])

# Compute similarity of a query vs indexed documents.
sims = index[query]
```



cnrs


projector.tensorflow.org
word embedding explorer - Recherc...
Embedding projector - visualization...
wevi
Vector explorers
Word embedding demo
Embedding Projector
DATA
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%.
Points: 10000 | Dimension: 200
Show All Data
Isolate 5948 points
Clear selection
Search
by
word
BOOKMARKS (0)
https://projector.tensorflow.org


projector.tensorflow.org
word embedding explorer - Recherc...
Embedding projector - visualization...
wevi
Vector explorers
Word embedding demo
Embedding Projector
DATA
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%.
Points: 10000 | Dimension: 200 | Selected 101 points
free
word free
count 5684
compact
software
smooth
download
simple
tools
content
easy
allows
available
virtual net
tool
license
licenses
basic
positive
source
etc
provides
specific
allowing
open here
online
process
development
free
creative
project
create
good
black
allow
required
freely
providing
which
new
fair
liberty
without
creation
and
benefits
thus
itself
non
but
about
friendly
support
independent
pro
restrictions
allowed
strong
freedom
union
Show All Data
Isolate 101 points
Clear selection
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

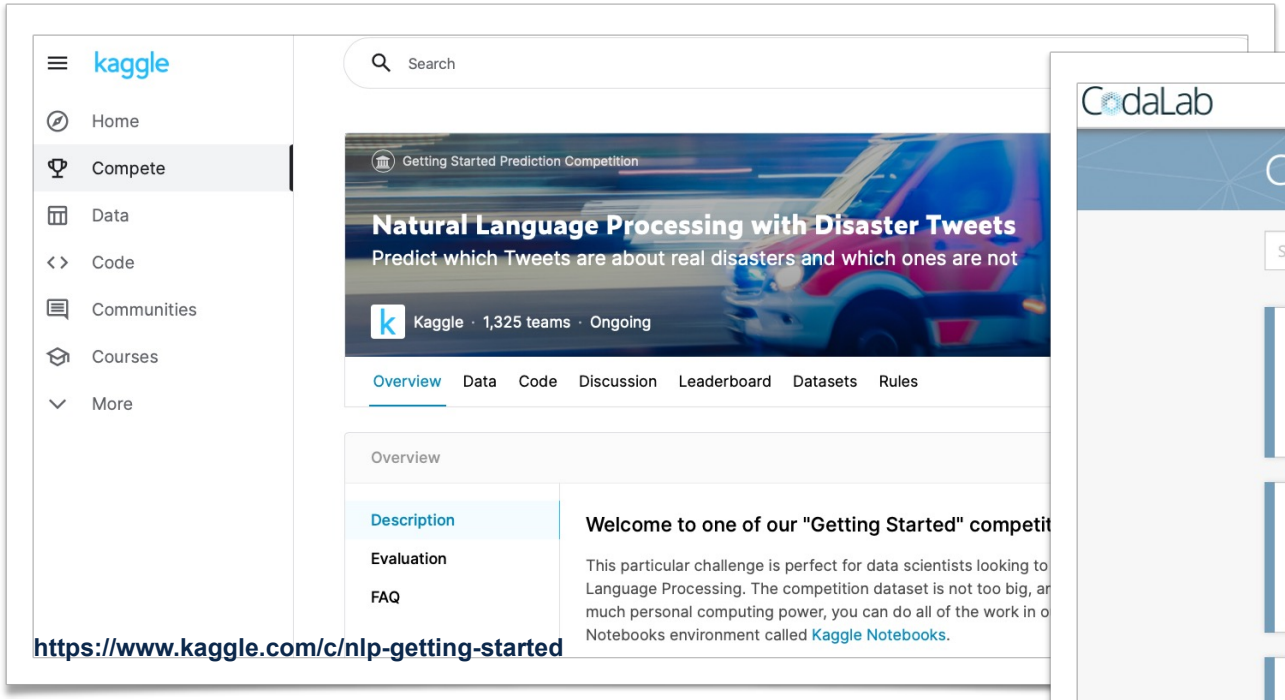
Social factors & opinion

- **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

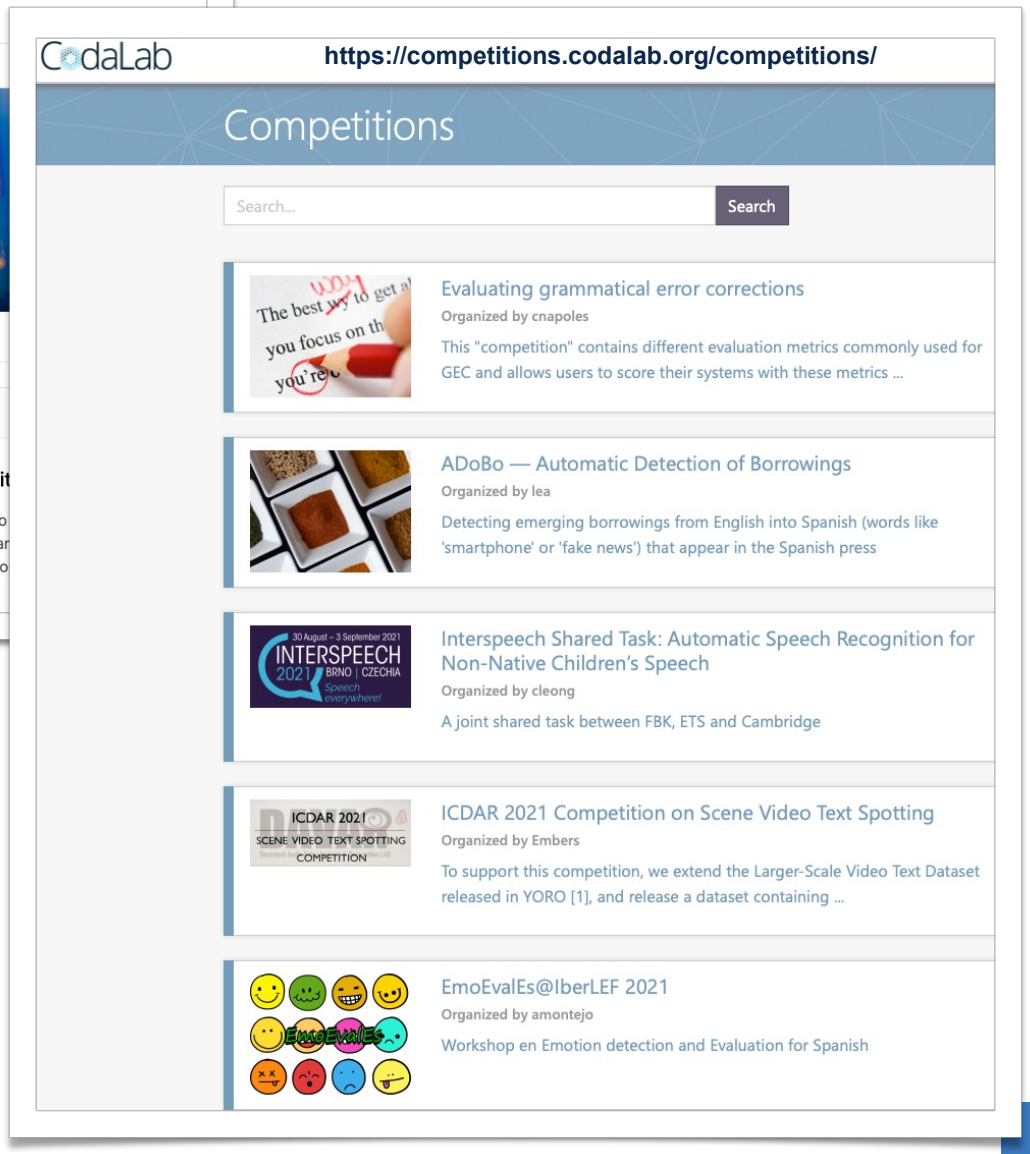
Information in scientific & clinical text

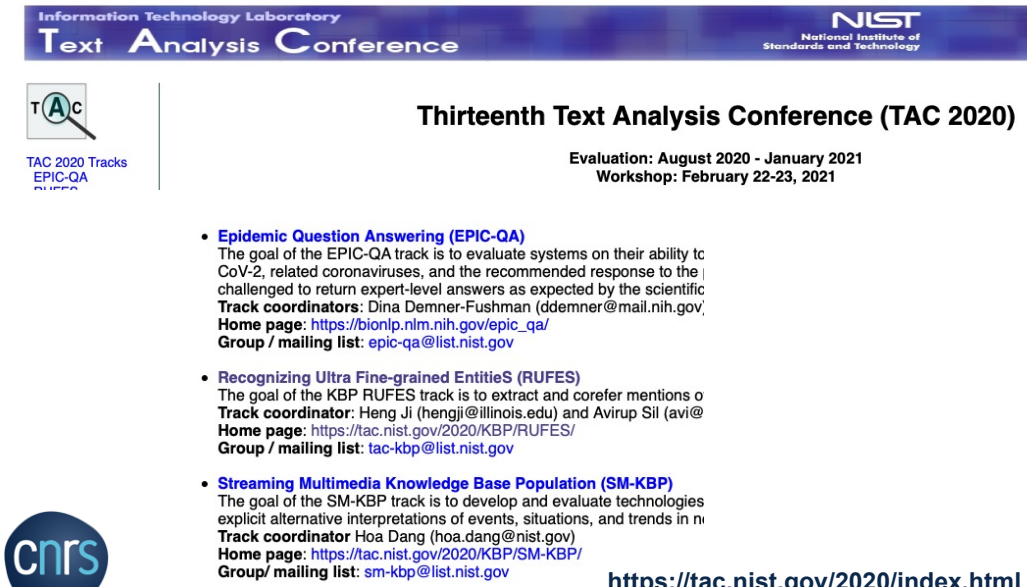
- **Task 8: MeasEval: Counts and Measurements** ([email organizers] [mailing list])
  Corey Harper, Jessica Cox, Ron Daniel, Paul Groth, Curt Kohler, Antony Scerri
- **Task 9: Statement Verification and Evidence Finding with Tables** ([email organizers] [mailing list])
  Nancy Xin Ru Wang, Sara Rosenthal, 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



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





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 :

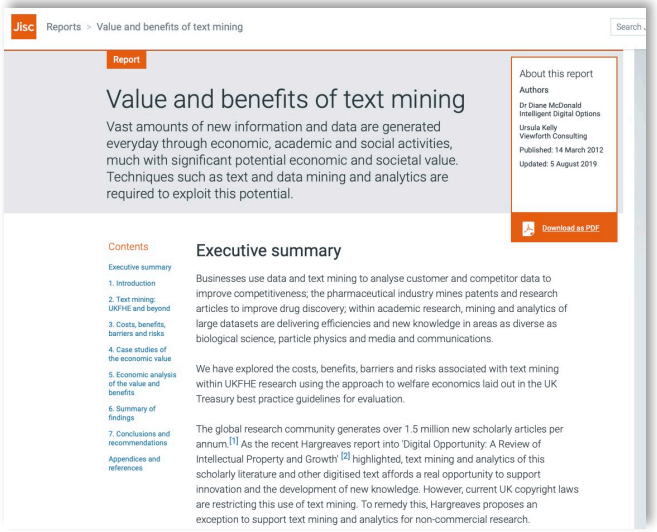
- un corpus cible, des ressources de spécialité
- d'intégrer différents composants logiciels
- un scénario et une référence pour évaluer la chaîne de traitements

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



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