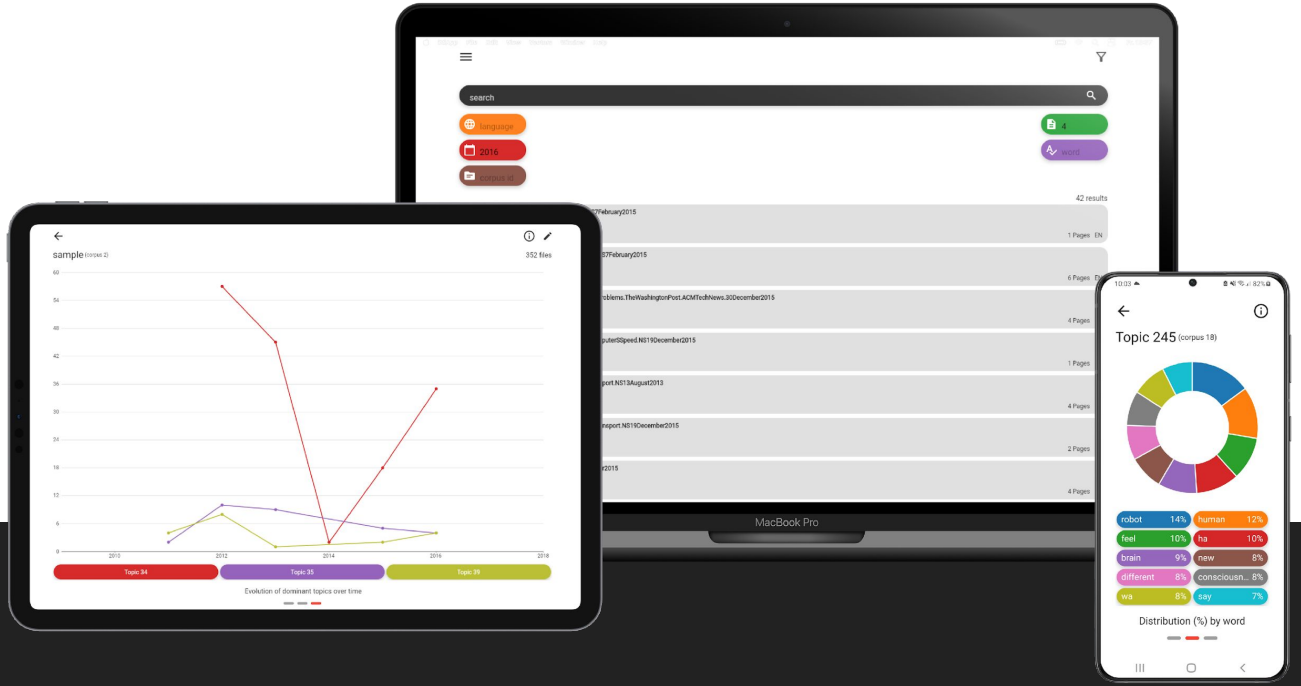
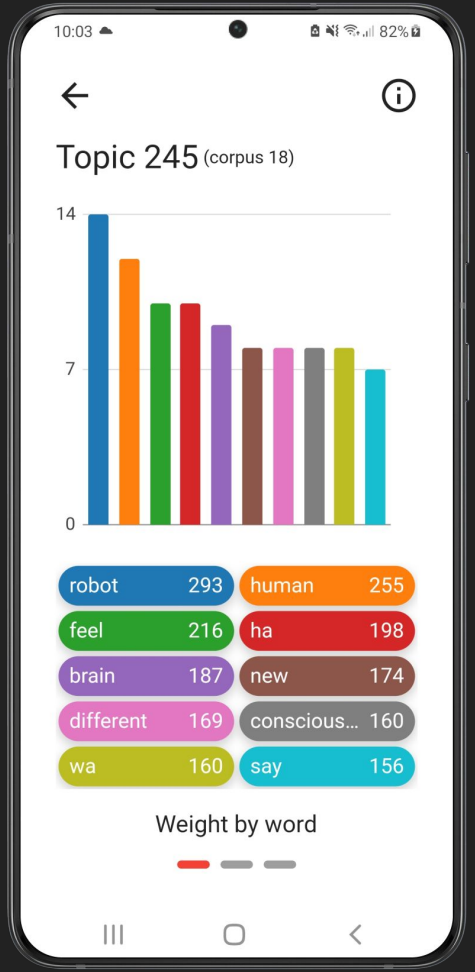


Visualisation d'une bibliographie personnelle automatiquement indexée et catégorisée

Projet innovant
2020-2021

Présenté par
Rémi Streignard



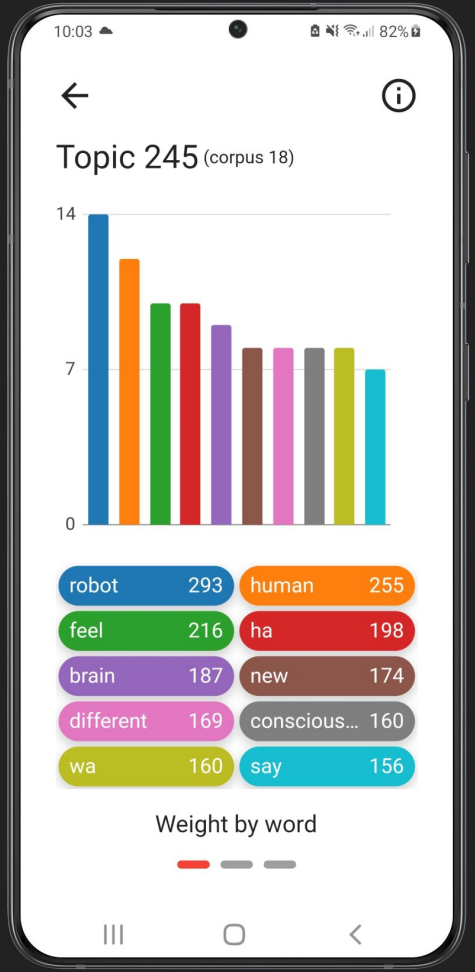
01 Problématique

02 Démarche

03 Observations

04 Analyse technique

05 Démonstration



01 Problématique

02 Démarche

03 Observations

04 Analyse technique

05 Démonstration

Problématique

part 1 : la découpe

Visualisation d’une bibliographie personnelle automatiquement indexée et catégorisée

01

visualisation

02

automatisation

03

indexation & catégorisation

Problématique

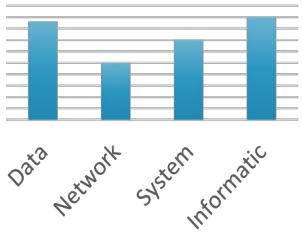
part 2 : les concepts

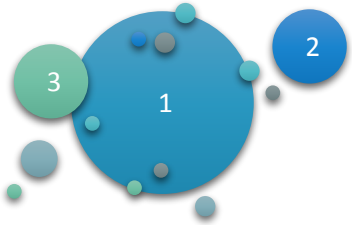
topic modeling

Modélisation de documents par l'utilisation d'algorithmes afin d'en déterminer les sujets abstraits.

corpus

topic



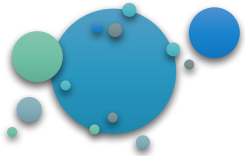


Problématique

part 3 : la vue simplifiée

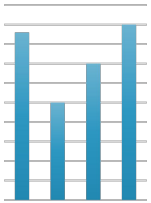
corpus

topic modeling

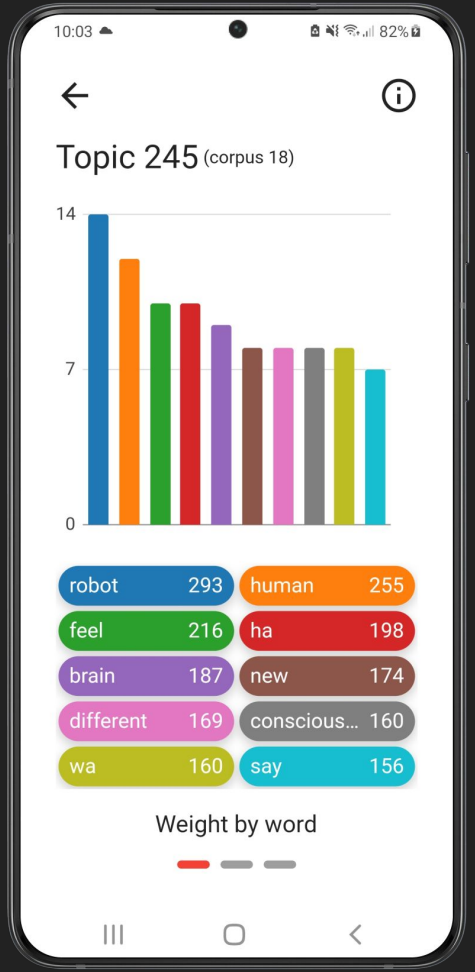


Dispersion des topics (dans un corpus)

Distribution des topics (dans un document)



Distribution des mots (dans un topic)



01 Problématique

02 Démarche

03 Observations

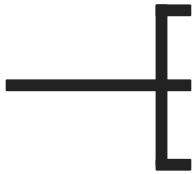
04 Analyse technique

05 Démonstration

Démarche

part 1 : le questionnement

Qu'évoque la problématique ?



Comment amener les données ? les rendre lisibles ?

Comment minimiser les actions de l'utilisateur ?

Que trier ? Grâce à quels attributs ?

Démarche

part 1 : le questionnement

Qu'évoque la problématique ?

Comment amener les données ? les rendre lisibles ?

Comment minimiser les actions de l'utilisateur ?

Que trier ? Grâce à quels attributs ?

Attentes de M. Philippe ?

Comment penser l'ergonomie ?

Quels visuels choisir ?

Comment éviter le "purement technique" ?

Démarche

part 1 : le questionnement

Qu'évoque la problématique ?

- Comment amener les données ? les rendre lisibles ?
- Comment minimiser les actions de l'utilisateur ?
- Que trier ? Grâce à quels attributs ?

Attentes de M. Philippe ?

- Comment penser l'ergonomie ?
- Quels visuels choisir ?
- Comment éviter le "purement mathématique" ?

Attentes du cours ?

- Qui est la cible ? Doit-elle changer ?
- Comment revisiter le concept ? Qu'apporter de plus ?

Démarche

part 1 : le questionnement

Qu'évoque la problématique ?

- Comment amener les données ? les rendre lisibles ?
- Comment minimiser les actions de l'utilisateur ?
- Que trier ? Grâce à quels attributs ?

Attentes de M. Philippe ?

- Comment penser l'ergonomie ?
- Quels visuels choisir ?
- Comment éviter le "purement mathématique" ?

Attentes du cours ?

- Qui est la cible ? Doit-elle changer ?
- Comment revisiter le concept ? Qu'apporter de plus ?

Démarche

part 2 : les valeurs

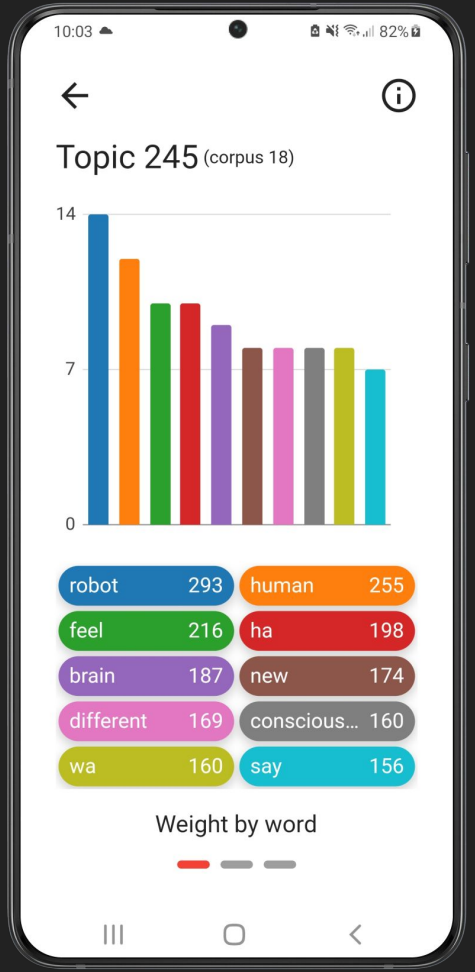
expérience

nouveauté

simplicité

lisibilité

sobriété



01 Problématique

02 Démarche

03 Observations

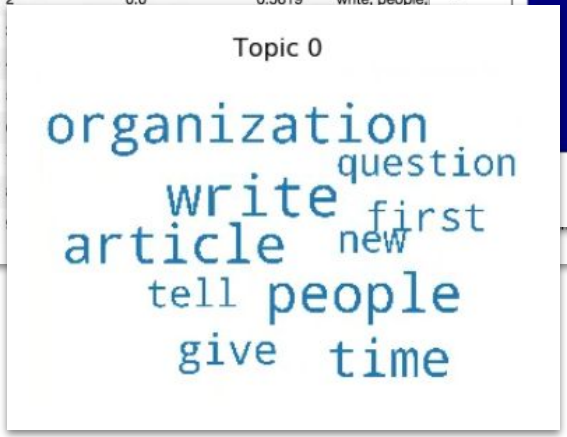
04 Analyse technique

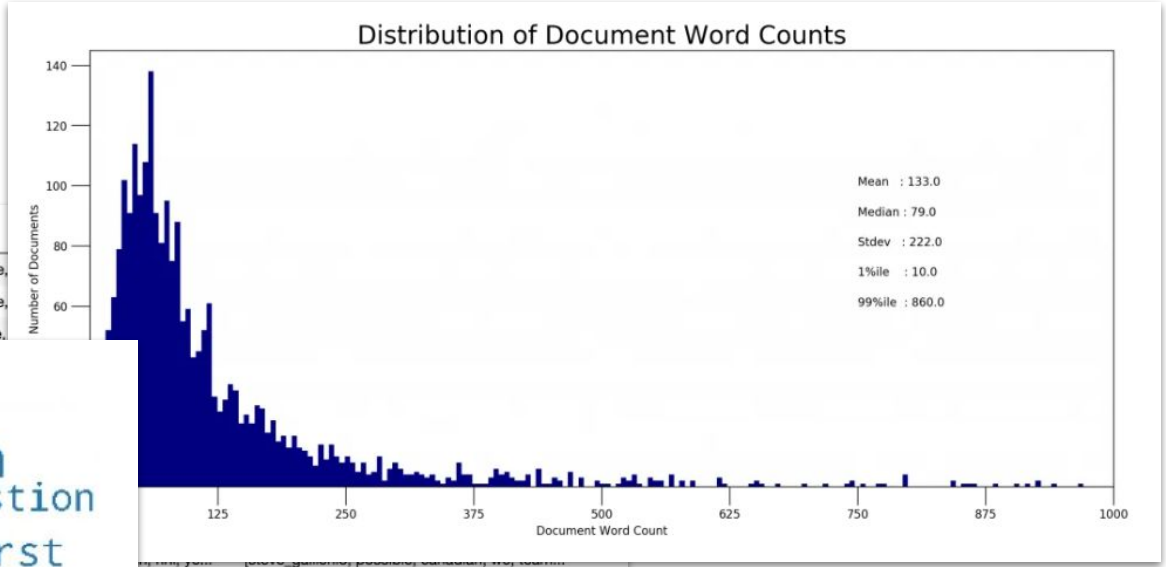
05 Démonstration

Observations

part 1 : les visuels

	Document_No	Dominant_Topic	Topic_Perc_Contrib	
0	0	2.0	0.6916	armenian, bike,
1	1	2.0	0.6247	armenian, bike,
2	2	0.0	0.5619	write, people,
3				
4				
5				
6				
7				
8				
9				

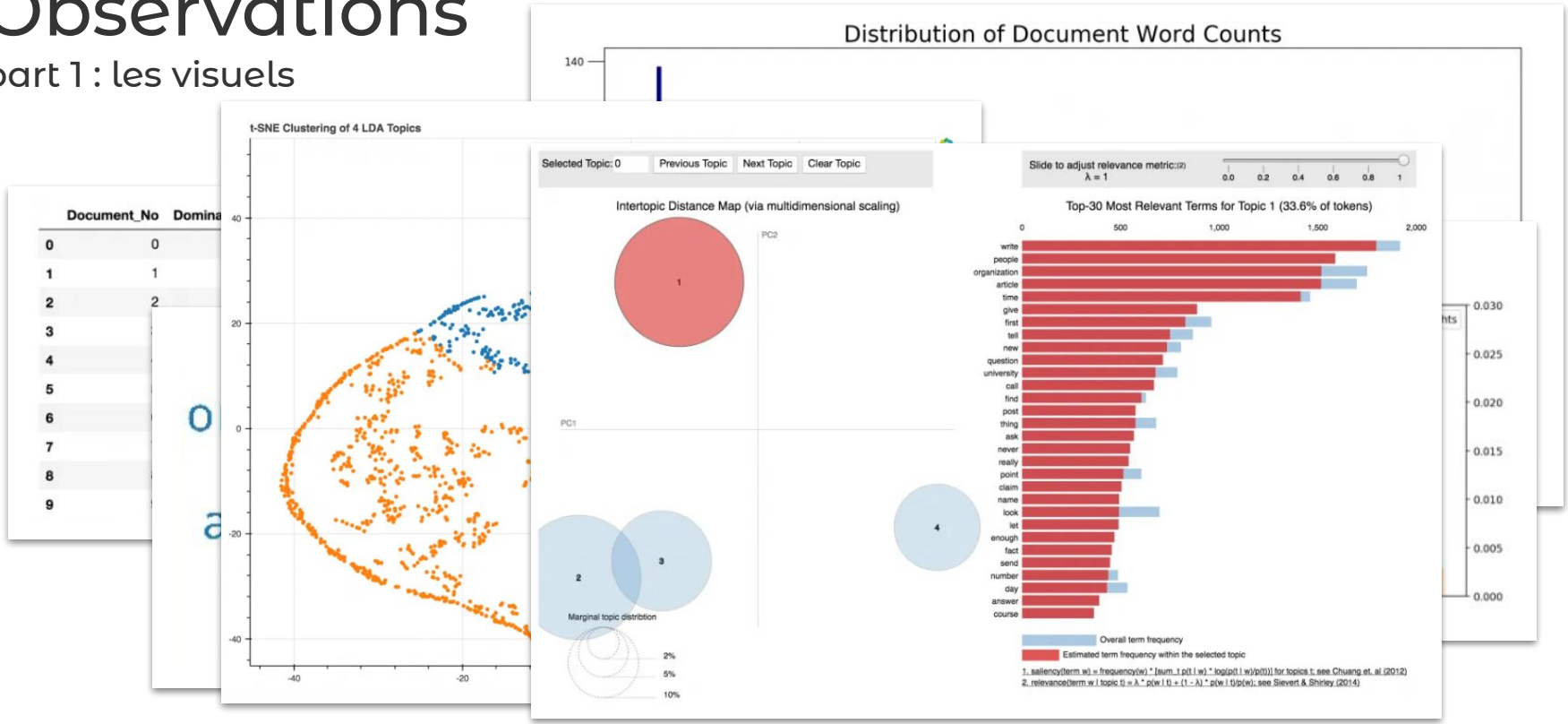




https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

Observations

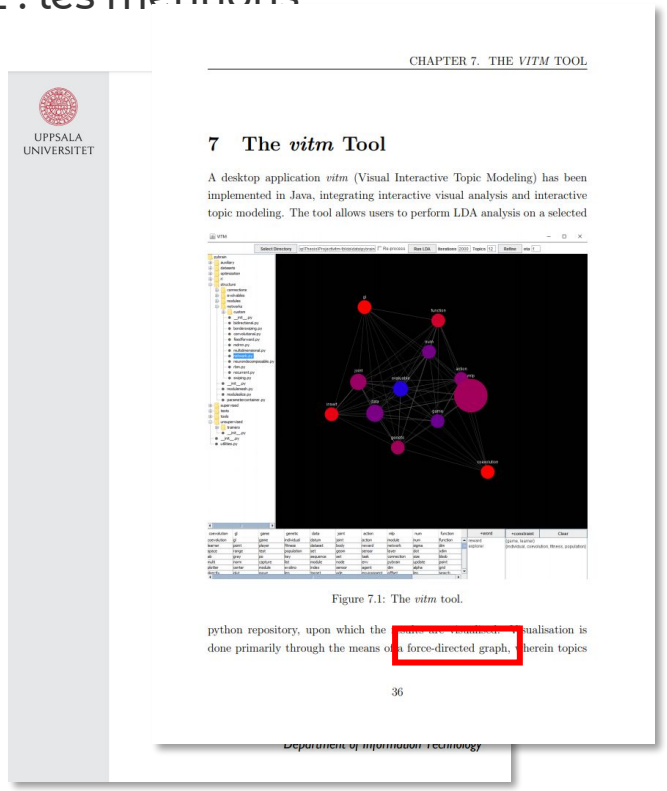
part 1 : les visuels



https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

Observations

part 2 : les mentions



UPPSALA UNIVERSITET

CHAPTER 7. THE *VITM* TOOL

7 The *vitm* Tool

A desktop application *vitm* (Visual Interactive Topic Modeling) has been implemented in Java, integrating interactive visual analysis and interactive topic modeling. The tool allows users to perform LDA analysis on a selected

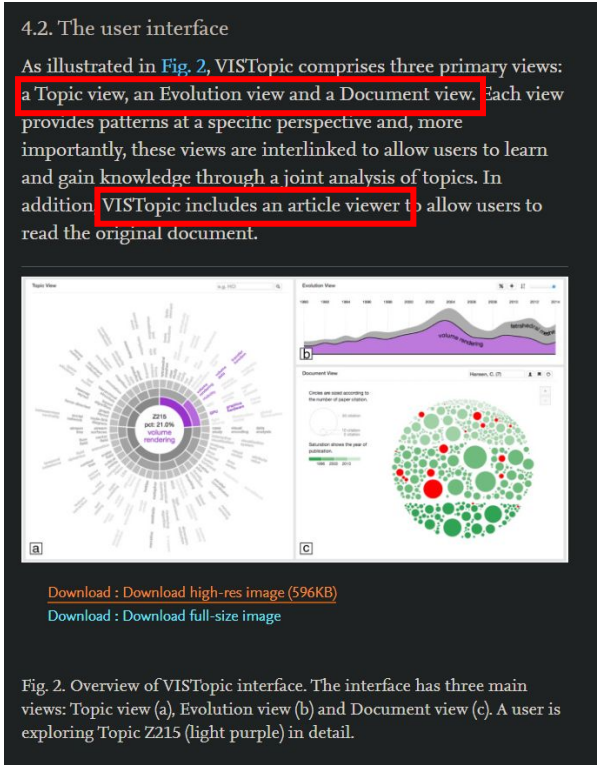
Figure 7.1: The *vitm* tool.

python repository, upon which the [redacted] visualisation is done primarily through the means of a force-directed graph, wherein topics

36

Department of Information Technology

https://uu.diva-portal.org/smash/get/diva2:1159462/FULLTEXT01.pdf



4.2. The user interface

As illustrated in Fig. 2, VISTopic comprises three primary views: a Topic view, an Evolution view and a Document view. Each view provides patterns at a specific perspective and, more importantly, these views are interlinked to allow users to learn and gain knowledge through a joint analysis of topics. In addition, VISTopic includes an article viewer to allow users to read the original document.

Download : Download high-res image (596KB)

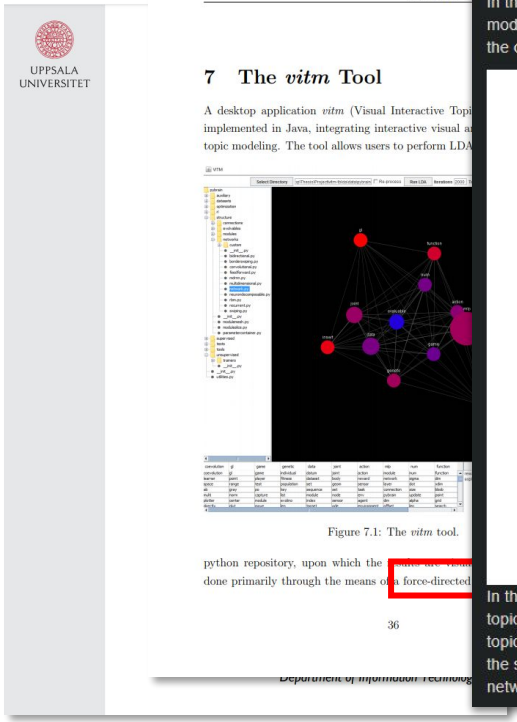
Download : Download full-size image

Fig. 2. Overview of VISTopic interface. The interface has three main views: Topic view (a), Evolution view (b) and Document view (c). A user is exploring Topic Z215 (light purple) in detail.

https://www.sciencedirect.com/science/article/pii/S2468502X17300074#fig2

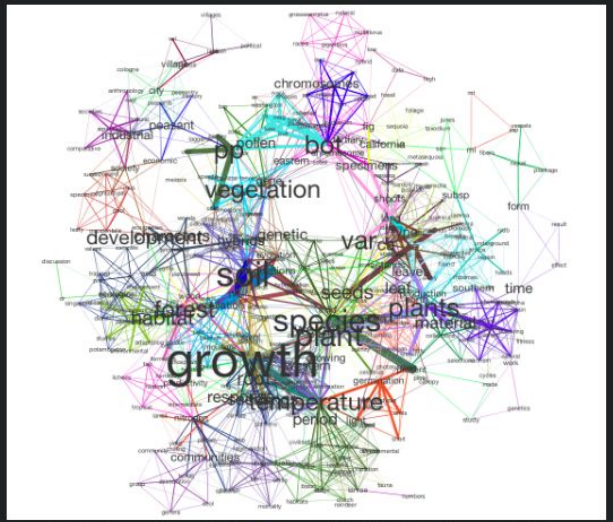
Observations

part 2 : les mentions



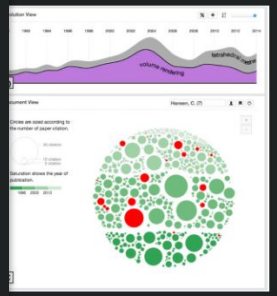
Tethne provides a variety of methods for working with text corpora and the output of modeling tools like MALLET. This tutorial focuses on parsing, modeling, and visualizing a Latent Dirichlet Allocation topic model, using data from the JSTOR Data-for-Research portal.

In this tutorial, we will use Tethne to prepare a JSTOR DfR corpus for topic modeling in MALLET, and then use the results to generate a semantic network like the one shown below.



In this visualization, words are connected if they are associated with the same topic; the heavier the edge, the more strongly those words are associated with that topic. Each topic is represented by a different color. The size of each word indicates the structural importance (betweenness centrality) of that word in the semantic network.

omprises three primary views:
d a Document view. Each view
spective and, more
inked to allow users to learn
nt analysis of topics. In
icle viewer to allow users to



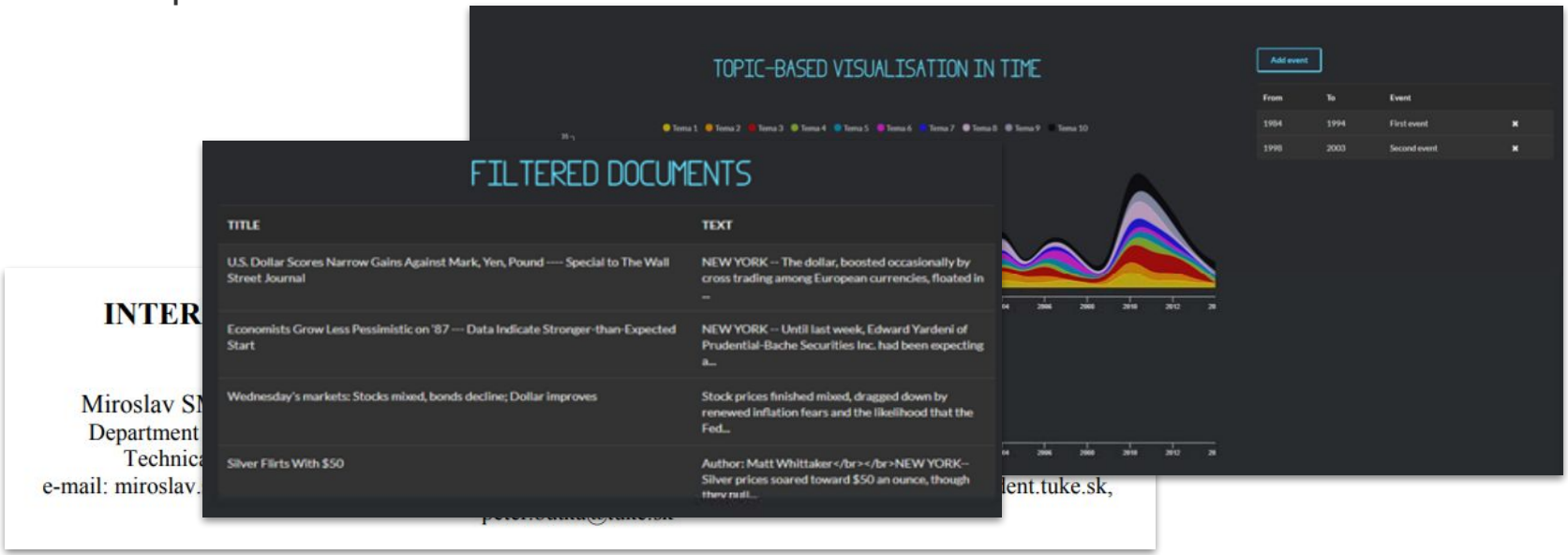
6KB)

The interface has three main
) and Document view (c). A user is
detail.

https://diging.github.io/tethne/api/tutorial.mallet.html

Observations

part 3 : le top 3



https://www.researchgate.net/publication/335263840_INTERACTIVE_TOOL_FOR_VISUALIZATION_OF_TOPIC_MODELS

Observations

part 3 : le top 3

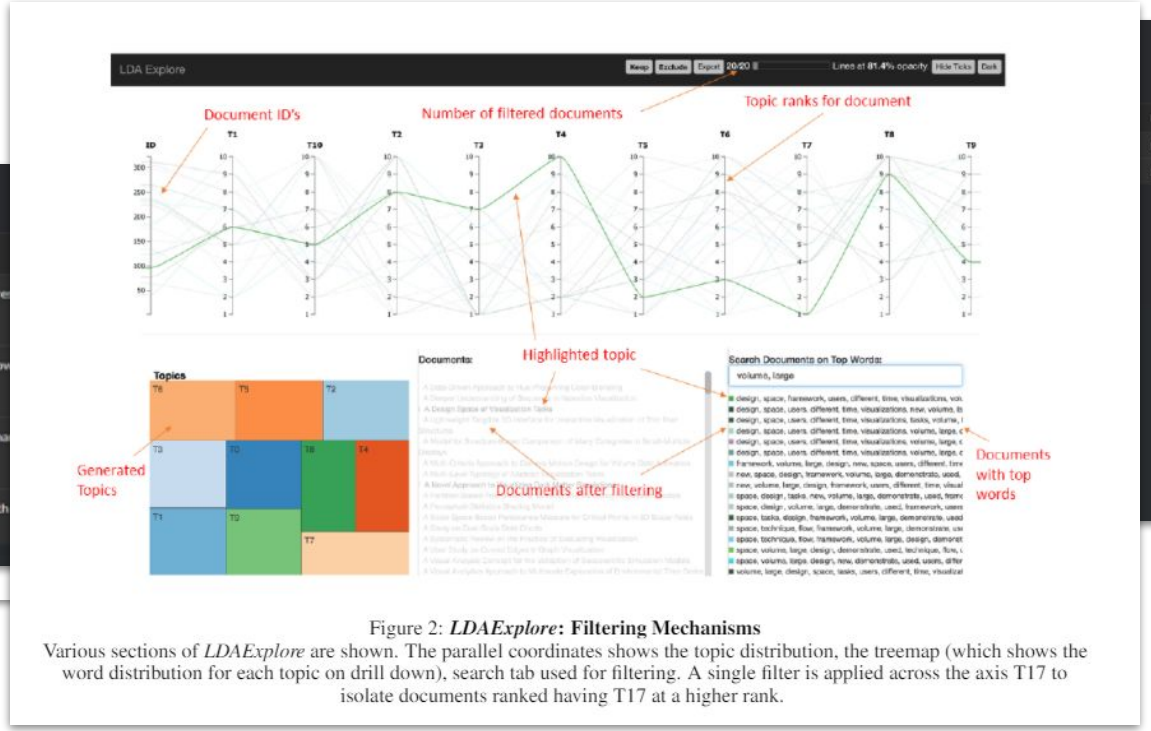


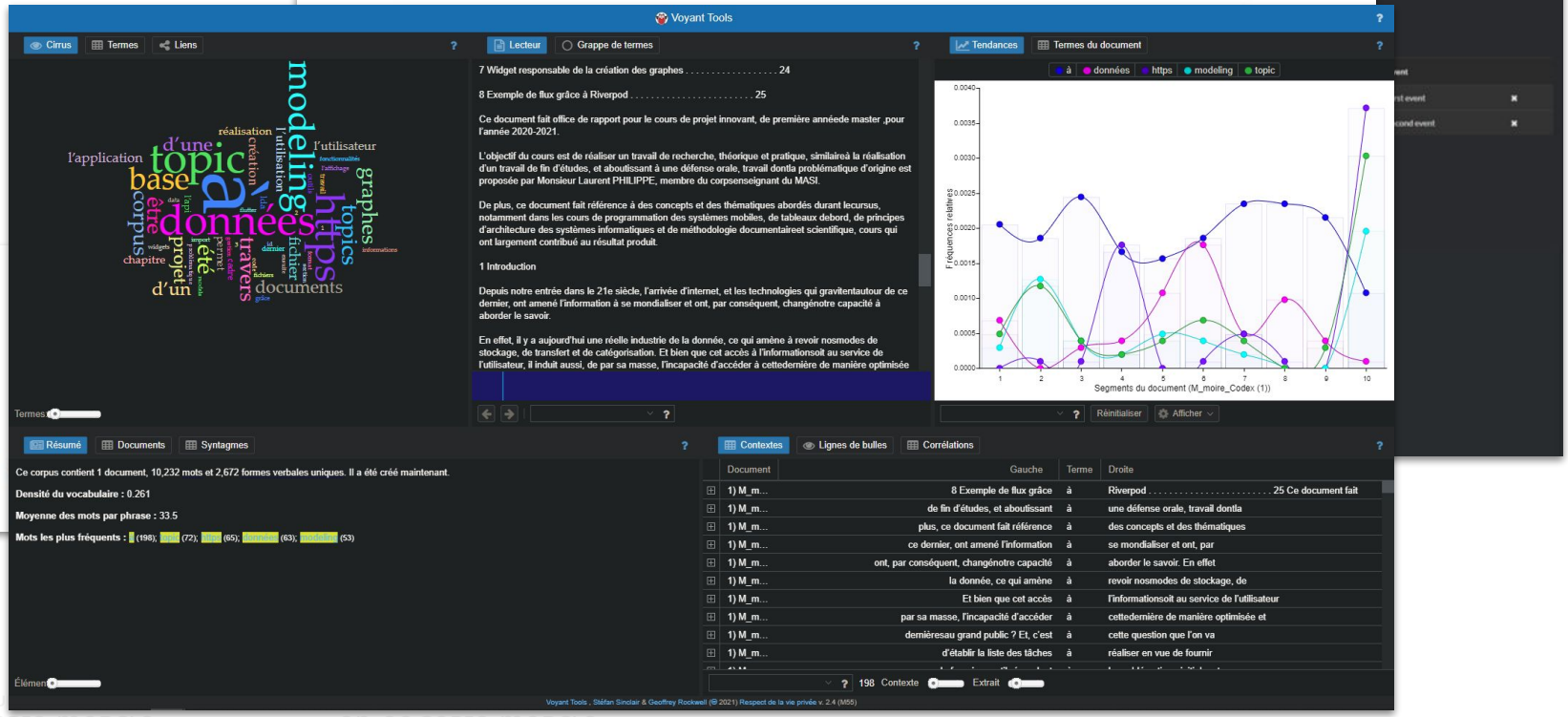
Figure 2: ***LDAExplore*: Filtering Mechanisms**
Various sections of *LDAExplore* are shown. The parallel coordinates shows the topic distribution, the treemap (which shows the word distribution for each topic on drill down), search tab used for filtering. A single filter is applied across the axis T17 to isolate documents ranked having T17 at a higher rank.

https://www.researchgate.net/publication/335263840_INTERACTIVE_TOOL_FOR_VISUALIZATION_OF_TOPIC_MODELS

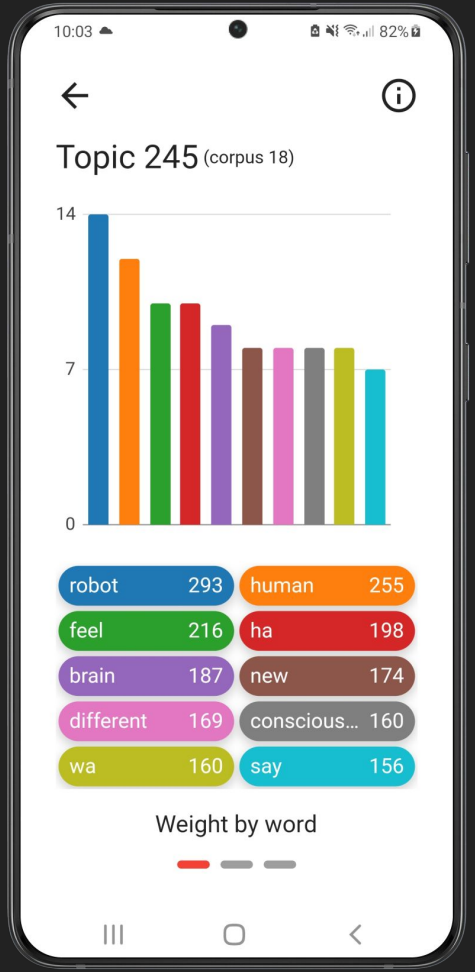
https://www.researchgate.net/publication/335263840_INTERACTIVE_TOOL_FOR_VISUALIZATION_OF_TOPIC_MODELS

Observations

part 3 : le top 3



https://voyant-tools.org/



01 Problématique

02 Démarche

03 Observations

04 Analyse technique

05 Démonstration

Analyse technique

part 1 : les entrées

	A	B	C	D	E	F	G	H	I	J	K
1		file_name	file_type	creation_time	file_size	n_pages	dt	error	language	without_stop_words	
2	1895	NotAllPossiblyRandomSeq	pdf	1134216656	316592	6	0.062504		EN	proc natl acad sci usa vol pp april ma	
3	1657	MatchingForensicSketches	pdf	1308301249	1396382	8	0.078123		EN	matching forensic sketch mug shot	
4	1130	GeneratingParticlelikeScatt	pdf	1308302379	1475594	4	0.046876		EN	generating particlelike scattering st	
5	397	bi2_white_paper	pdf	1308666006	1088104	16	0.140661		EN	business integrated insight bi reinv	
6	807	DiscreteWaveletTransform	pdf	1308856705	795492	37	0.171879		EN	discrete wavelet transform based t	
7	1584	LockSmithPracticalStaticRa	pdf	1308856767	1932873	55	0.156245		EN	locksmith practical static race detec	
8	3098	WYSINWYXWhatYouSeeIsW	pdf	1308856963	1848031	84	0.249971		EN	wysinwyx see execute gogul balakr	
9	215	AnomalyDetectionASurvey	pdf	1308857047	725548	58	0.28362		EN	anomaly detection survey varun ch	
10	1643	MakingTheDevelopmentCl	pdf	1309459205	669655	8	0.031246		EN	june making development cloud rea	
11	1	110704_NOTE DE BASE FORI	pdf	1309803609	1723472	113	0.33214		FR	etat federal plus efficace entites pl	
12	2119	ProtocolsForBuildingAnOrg	pdf	1317672368	4548488	24	0.078123		EN	int j knowledge learning vol copyrig	
13	475	BuildingTheEncoreDictiona	pdf	1317672513	1747355	14	0.182188		EN	int j knowledge learning vol copyrig	
14	2862	TwitterMoodPredictsTheSt	pdf	1323279363	357610	8	0.078132		EN	twitter mood predicts stock market	
15	1528	LearningActivityPatternsUs	pdf	1323280475	770206	9	0.078128		EN	ieee transaction system man cyber	
16	2861	TwitterMoodAsAStockMark	pdf	1323280788	2132006	4	0.062493		EN	october discovery analytics publish	
17	1554	LeonardosRuleSelfSimilarit	pdf	1327334052	450990	5	0.062501		EN	leonardo rule self similarity wind in	
18	2237	RealizingAllSpinBasedLogic	pdf	1327605207	587417	5	0.062498		EN	doi science et al alexander ako khaj	
19	32	AccurateMeasurementOfC	pdf	1328294701	2674743	10	0.203125		EN	eur phys j appl phys doi european p	
20	1815	NearlyOptimalSparseFouri	pdf	1328295408	254011	27	0.125007		EN	arxiv jan nearly optimal sparse fouri	
21	2064	PowerDesigner_16.0_Read	pdf	1328560463	38257	1	0.015627		EN	dc sybase powerdesigner read first	
22	615	ComputingPerformanceGai	pdf	1329165412	1128516	8	0.125047		EN	cover feature january published iee	

Fichier au format CSV exporté au format excel

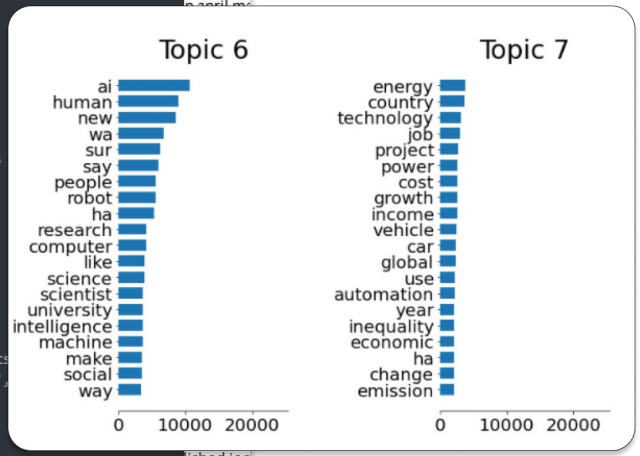
Analyse technique

part 1 : les entrées

	A	B	C	D
1		file_name	file_type	creation_time
2	1895	NotAllPossiblyRandomSeq	pdf	113421665
3	1657	MatchingForensicSketches	pdf	130830124
4	1130	GeneratingParticlelikeScatt	pdf	130830237
5	397	bi2_white_paper	pdf	130866600
6	807	DiscreteWaveletTransform	pdf	130885670
7	1584	LockSmithPracticalStaticRa	pdf	130885676
8	3098	WYSINWYXWhatYouSeeIsW	pdf	130885696
9	215	AnomalyDetectionASurvey	pdf	130885704
10	1643	MakingTheDevelopmentCl	pdf	130945920
11	1	110704_NOTE DE BASE FORI	pdf	130980360
12	2119	ProtocolsForBuildingAnOrg	pdf	131767236
13	475	BuildingTheEncoreDictiona	pdf	131767251
14	2862	TwitterMoodPredictsTheSt	pdf	132327936
15	1528	LearningActivityPatternsUs	pdf	132328047
16	2861	TwitterMoodAsAStockMark	pdf	132328078
17	1554	LeonardosRuleSelfSimilarit	pdf	132733405
18	2237	RealizingAllSpinBasedLogic	pdf	132760520
19	32	AccurateMeasurementOfC	pdf	132829470
20	1815	NearlyOptimalSparseFouri	pdf	132829540
21	2064	PowerDesigner_16.0_Read	pdf	132856046
22	615	ComputingPerformanceGa	pdf	132916541

Fichier au format CSV exporté

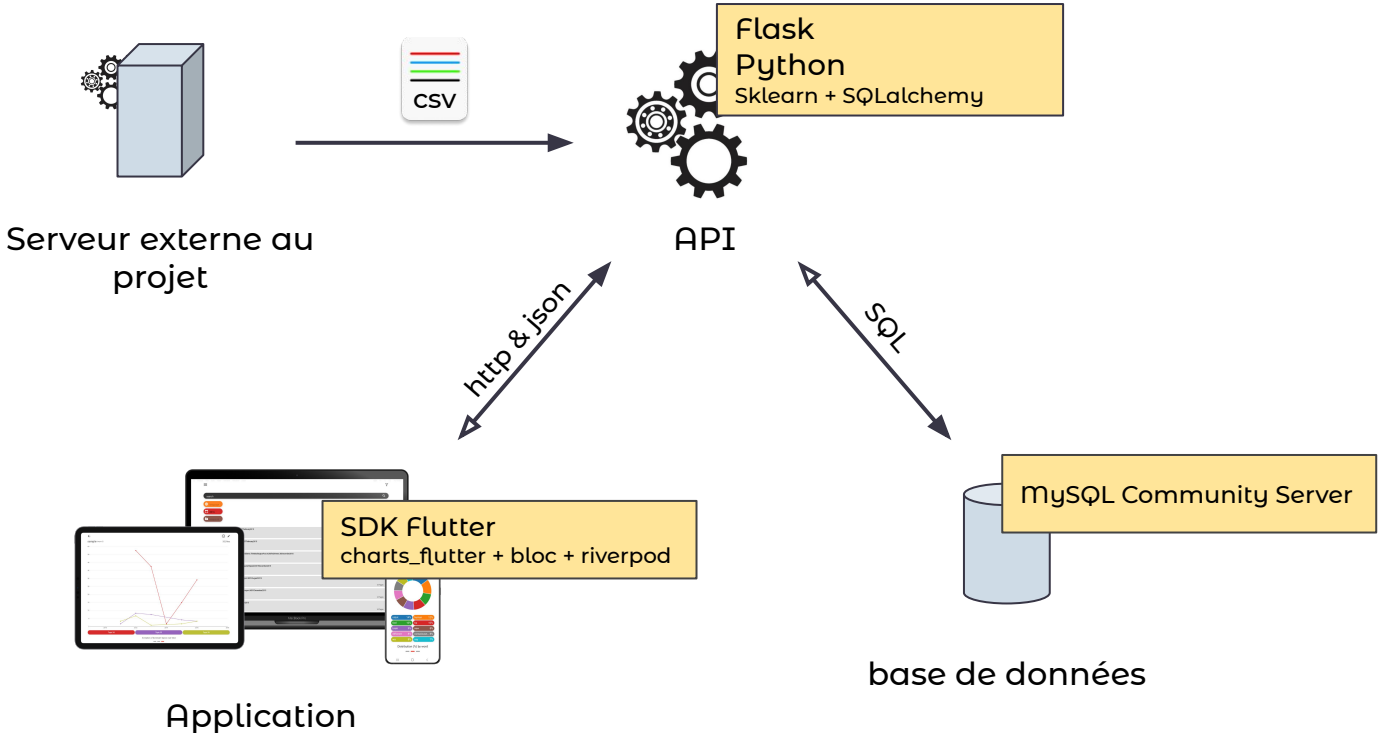
```python
# Use tf-idf features for NMF.
print("Extracting tf-idf features")
tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2,
                                   max_features=n_features,
                                   stop_words='english')
t0 = time()
tfidf = tfidf_vectorizer.fit_transform(data_samples)
print("done in %0.3fs." % (time() - t0))

# Use tf (raw term count) features for LDA.
print("Extracting tf features for LDA...")
tf_vectorizer = CountVectorizer(max_df=0.95, min_df=2,
                                max_features=n_features,
                                stop_words='english')
t0 = time()
tf = tf_vectorizer.fit_transform(data_samples)
print("done in %0.3fs." % (time() - t0))
print()

print('\n' * 2, "Fitting LDA models with tf features, "
      "n_samples=%d and n_features=%d..."
      % (n_samples, n_features))
lda = LatentDirichletAllocation(n_components=n_components,
                                learning_method='online',
                                learning_offset=50.,
                                random_state=0)
t0 = time()
lda.fit(tf)
print("done in %0.3fs." % (time() - t0))

tf_feature_names = tf_vectorizer.get_feature_names()
plot_top_words(lda, tf_feature_names, n_top_words, 'Topics in LDA model')
```



Fichier notebook (Python) responsable du topic modeling

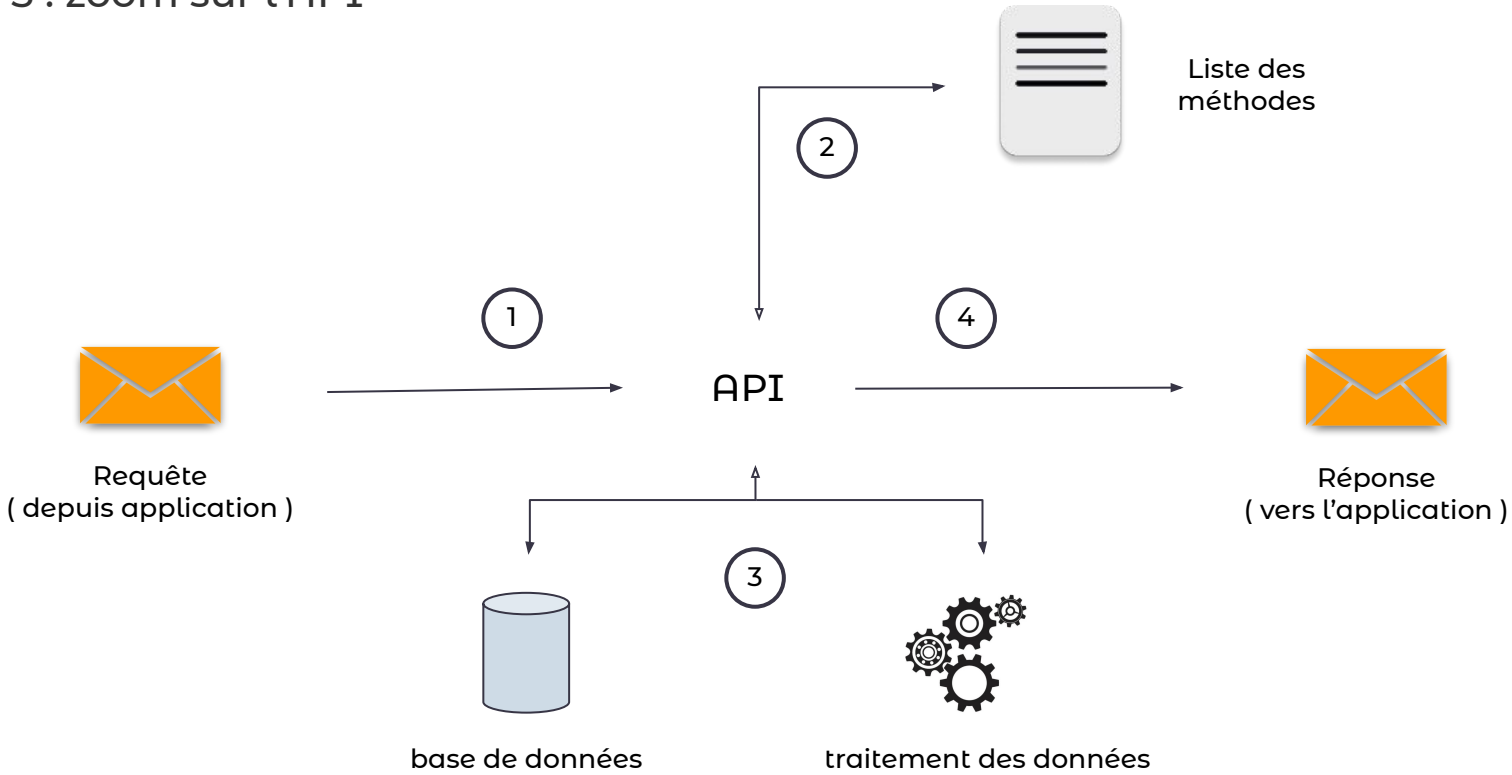
Analyse technique

part 2 : l'architecture



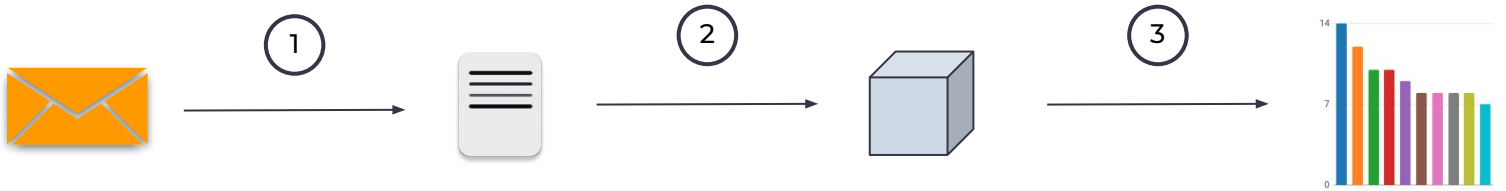
Analyse

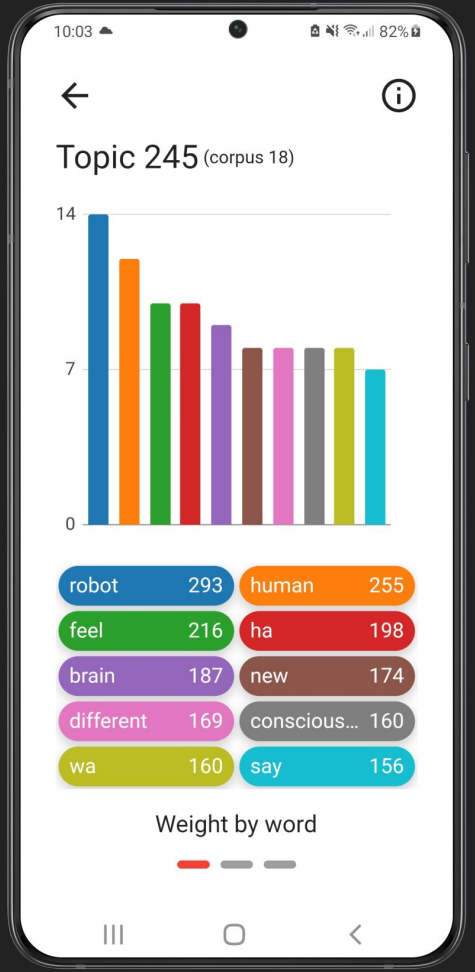
part 3 : zoom sur l'API



Analyse technique

part 4 : zoom sur l'application





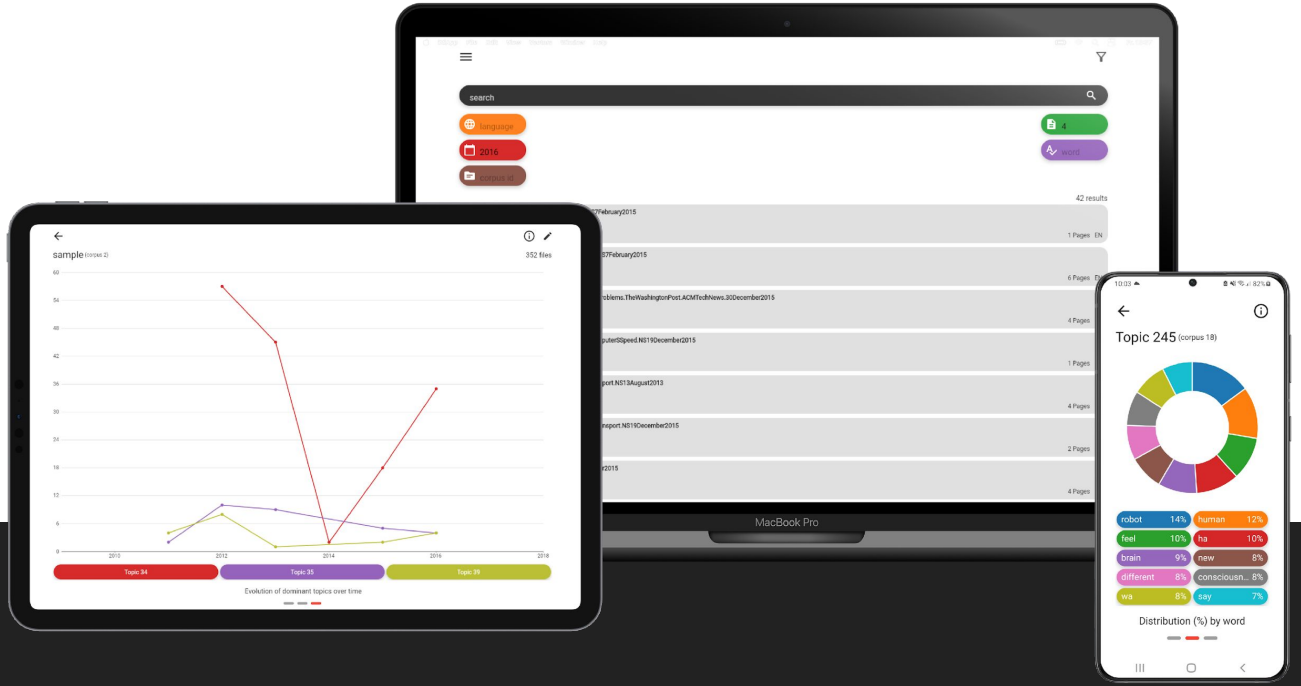
01 Problématique

02 Démarche

03 Observations

04 Analyse technique

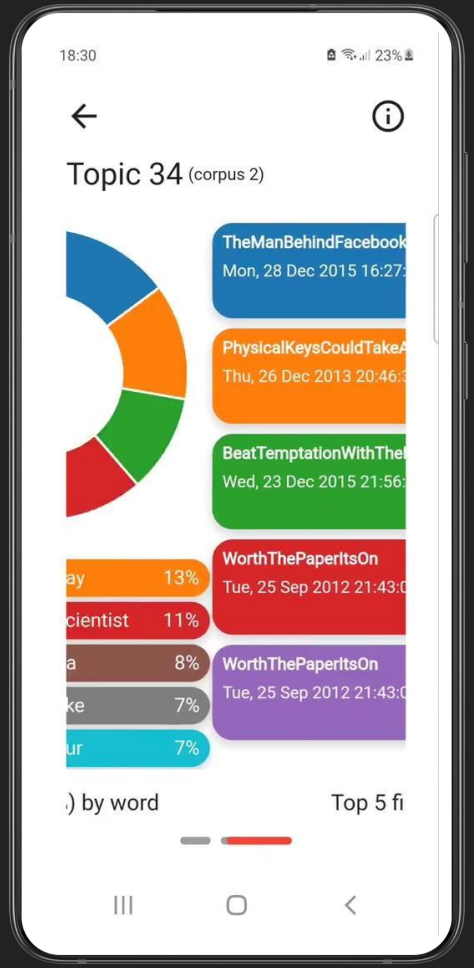
05 Démonstration



Merci de votre écoute !
Des questions ?

Projet innovant
2020-2021

Présenté par
Rémi Streignard


18:30
23%
Topic 34 (corpus 2)
TheManBehindFacebook
Mon, 28 Dec 2015 16:27
PhysicalKeysCouldTakeA
Thu, 26 Dec 2013 20:46
BeatTemptationWithThe
Wed, 23 Dec 2015 21:56
WorthThePaperItsOn
Tue, 25 Sep 2012 21:43:0
13%
cientist 11%
8%
WorthThePaperItsOn
Tue, 25 Sep 2012 21:43:0
7%
7%
) by word
Top 5 fi

Plan B