Collecte des Données (1/2)

Problèmes:

- Pas de corpus a priori
- Pas d'API
- Pas de structuration uniforme
- Pas d'encodage uniforme
- Pas d'unicité des références

Collecte des Données (2/2)

Démarche:

- Identification arbitraire d'un corpus avec temps d'arrêt
- Extraction récursive des liens avec élagage
- Extraction des contenus
- Extraction des textes propres
- Décodage
- Stopwords et lemmatisation

```
http://www.profmichelle.com/2017/01/nos-cahiers-de-noel.htm
```

```
http://www.profmichelle.com/2017/01/nos-cahiers-de-noel.html
```

```
"color:#000080;"><span lang="FR" style="background:white">Auschwitz</span>
"FR"><span style="background:white">.</span></span><span lang="FR"> L&#39;
%agrave: la 2<sup>e</sup> place du classement des 50 BD essentielles €eact
</i></span></span></span></span></span></div>
<div class="ob-section ob-section-link"><div class="ob-ctn ob-ctn-withIma"</pre>
https://www.kizoa.it/Movie-Maker/d92418675kP302368765o111/la-shoa-et-la-ba
ob-pull-left.ob-media-left" target=" blank"><img.src="
http://resize.over-blog.com/170x170-ct.jpg?http://www.kizoa.com/ajax/crops
class="ob-media ob-img" alt=
"Movie&#x20:Maker&#x3A:&#x20:La&#x20:Shoa&#x20:et&#x20:la&#x20:bande&#x20:
"ob-title"><a href="https://www.kizoa.it/Movie-Maker/d92418675kP302368765c
"ob-link" target=" blank">Movie Maker: La Shoa et la bande dessinée Maus</
non è stata attivata dalla persona che ha fatto questa creazione. Puoi ric
personalizzato. Questa funzionalità è disponibile ...
https://www.kizoa.it/Movie-Maker/d92418675kP3023687650111/la-shoa-et-la-ba
https://www.kizoa.it/Movie-Maker/d92418675kP302368765o111/la-shoa-et-la-ba
            <div class="ob-section ob-section-html">style="margin-botto">style="margin-botto">style="margin-botto"
sont les animaux qui apparaissent le plus fré quemment dans les imac
```

vocabulaire planche entiere Bande Dessinee compose cases vig vignettes indiquer changement lieu temps appeles elements na phylactere prend forme bulle contient paroles forme nuage co personnage holocauste Maus Spiegelman Maus bande dessinee ec Unis York traitant persecutions juifs annees 1930 1940 Shoah univers animalier nazis representes chats Juifs souris Polon oeuvre recu prix Pulitzer special 1992 traduite huit langues narratives pere Spiegelman racontant deportation recit relat raconte Vladek Spiegelman rescape juif camps nazis fils aute pere portes Auschwitz trottoirs York deroule temps annees 19 survivre survivant Spiegelman rend appartement pere York Regensemble evenements vecus pere lors Seconde Guerre mondiale premieres pages Spiegelman aborde theme majeur litterature SI solidarite deportes pere jeune fils pleure amis partis amis amis argent sauver selectionnes Auschwitz album classe 2012 p

Topic model (1/2)



chaque topic est une distribution de mots

Topic model (1/2)



chaque **topic** est une distribution de mots chaque **document** est un mélange de quelques topics

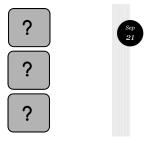
Topic model (1/2)



chaque **topic** est une distribution de mots chaque **document** est un mélange de quelques topics

chaque mot est tiré au sort dans un topic

Topic Model (2/2)



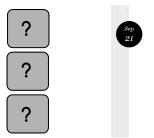
Foutage de gueule

Il paraît que le prof doc n'évalue pas les élèves.
Le prof doc passe un Capes! Certificat d'Aptitude au
Professorat de l'Enseignement du Second degré! Comment
peut-on enseigner sans évaluer? Sans faire un suivi des
élèves? Là, j'aimerais bien qu'on m'explique....
J'enseigne et j'évalue mes élèves. Je suis prof, c'est mon
boulot, faire passer des connaissances et m'assurer qu'elles
sont comprises, acquises.

On l'entend partout, notre ministre le dit partout,

Dans la réalité, on observe les documents

Topic Model (2/2)



Foutage de gueule

Il paraît que le prof doc n'évalue pas les élèves.
Le prof doc passe un Capes! Certificat d'Aptitude au
Professorat de l'Enseignement du Second degré! Comment
peut-on enseigner sans évaluer? Sans faire un suivi des
élèves? Là, j'aimerais bien qu'on m'explique....
J'enseigne et j'évalue mes élèves. Je suis prof, c'est mon
boulot, faire passer des connaissances et m'assurer qu'elles
sont comprises. acquises.

On l'entend partout, notre ministre le dit partout,

Dans la réalité, on observe les documents Tout le reste constitue des **variables cachées** Nous cherchons à les retrouver, en inversant le processus génératif

6 → liberte etat societe homme droit politique droi economie selon monde pouvoir liberalisme justice se 7 → triangle loys puzzle rectangle puzzles eleves z maths dernier mathix emes reponses exemple isocele 8 → documents republique document trace travail cor place entreprise etats france problematique composi 9 → jeux console jouer generation playstation stree attends dois couleurs nouvelle premier semble live 10 → eleves histoire premiere travail etaient ulysse terre matiere niveau sujets possible princesse troi 11 → opinion verite lire challenge opinions albums 12 → ecole eleves dalton classe acte parents educati

```
le-data\bigbag\l-laviemoderne-1.txt 0 0 association 73
le-data\bigbag\l-laviemoderne-1.txt 1 1 populaire 84
le-data\bigbag\l-laviemoderne-1.txt 2 2 victime 5
le-data\bigbag\l-laviemoderne-1.txt 3 3 numerisme 82
le-data\bigbag\l-laviemoderne-1.txt 4 4 etude 82
le-data\bigbag\l-laviemoderne-1.txt 5 5 instructive 1
le-data\bigbag\l-laviemoderne-1.txt 6 6 enseignements 71
le-data\bigbag\l-laviemoderne-1.txt 7 7 soient 57
le-data\bigbag\l-laviemoderne-1.txt 8 8 scientifiques 1
le-data\bigbag\l-laviemoderne-1.txt 9 9 nouveaux 82
le-data\bigbag\l-laviemoderne-1.txt 10 10 vient 5
le-data\bigbag\l-laviemoderne-1.txt 11 11 publier 84
le-data\bigbag\l-laviemoderne-1.txt 12 12 afev 82
le-data\bigbag\l-laviemoderne-1.txt 13 13 numerique 82
le-data\bigbag\l-laviemoderne-1.txt 14 14 collegiens 1
```

Limites et espoirs

- Hypothèse forte du modèle
- Gérer l'hétérogénéité (sur le fond) du corpus
- Démarche en échantillonnage/test pour dépasser l'exploratoire.
- Vers une constitution itérative du corpus