#HackaTAL2017

Tâches Brevet : exploration du corpus complet
https://github.com/nicolasdugue/hackatal2017

Marion Bernard, Kévin Deturcq, Nicolas Dugué, Loïc Grobol, Nadège Lechevrel, Frédéric Moal

Karaoké Team

Table of contents

Introduction

Tâche brevets chez la Team Karaoké

Objectif : Explorer l'**intégralité** du corpus.

Tâche brevets chez la Team Karaoké

Objectif : Explorer l'**intégralité** du corpus.

- Nécéssite
  - beaucoup de pré-traitements : définir **vocabulaire** ;
  - des algorithmes de **faible complexité** ;
  - des moyens d'**interpréter** les résultats.

Tâche brevets chez la Team Karaoké

Objectif : Explorer l'**intégralité** du corpus.

- Nécéssite
  - beaucoup de pré-traitements : définir **vocabulaire** ;
  - des algorithmes de **faible complexité** ;
  - des moyens d'**interpréter** les résultats.
- Pistes d'exploration, exploiter
  - l'axe temporel ;
  - l'axe catégoriel (A,B, C, ..., H) ;
  - les distances entre les mots du vocabulaire sur ces axes ;
  - les similarités sémantiques entre les mots du vocabulaire ;
  - la spécificité/représentativité des termes selon les années/catégories.

Pré-traitements

- Lemmatisation ;
- Suppression des nombres ;
- Suppression des hapax ;
- Possibilité de filtrer le vocabulaire :
    - selon le nombre de documents minimum dans lequel le lemme apparait ;
    - selon le nombre de documuments maximum ;
    - selon le nombre de catégories maximum.

Explorer les co-évolutions 1/2

Distances temporelles et catégorielles

Détecter les mots les plus proches d'un mot donné selon :

- l'histogramme temporel ;
- l'histogramme catégoriel.

Distances temporelles et catégorielles

Détecter les mots les plus proches d'un mot donné selon :

- l'histogramme temporel ;
- l'histogramme catégoriel.

Créer une distance :

- Kullback-Leibler/euclidienne sur les deux histogrammes ;
- prendre le max.

Distances temporelles et catégorielles

Détecter les mots les plus proches d'un mot donné selon :

- l'histogramme temporel ;
- l'histogramme catégoriel.

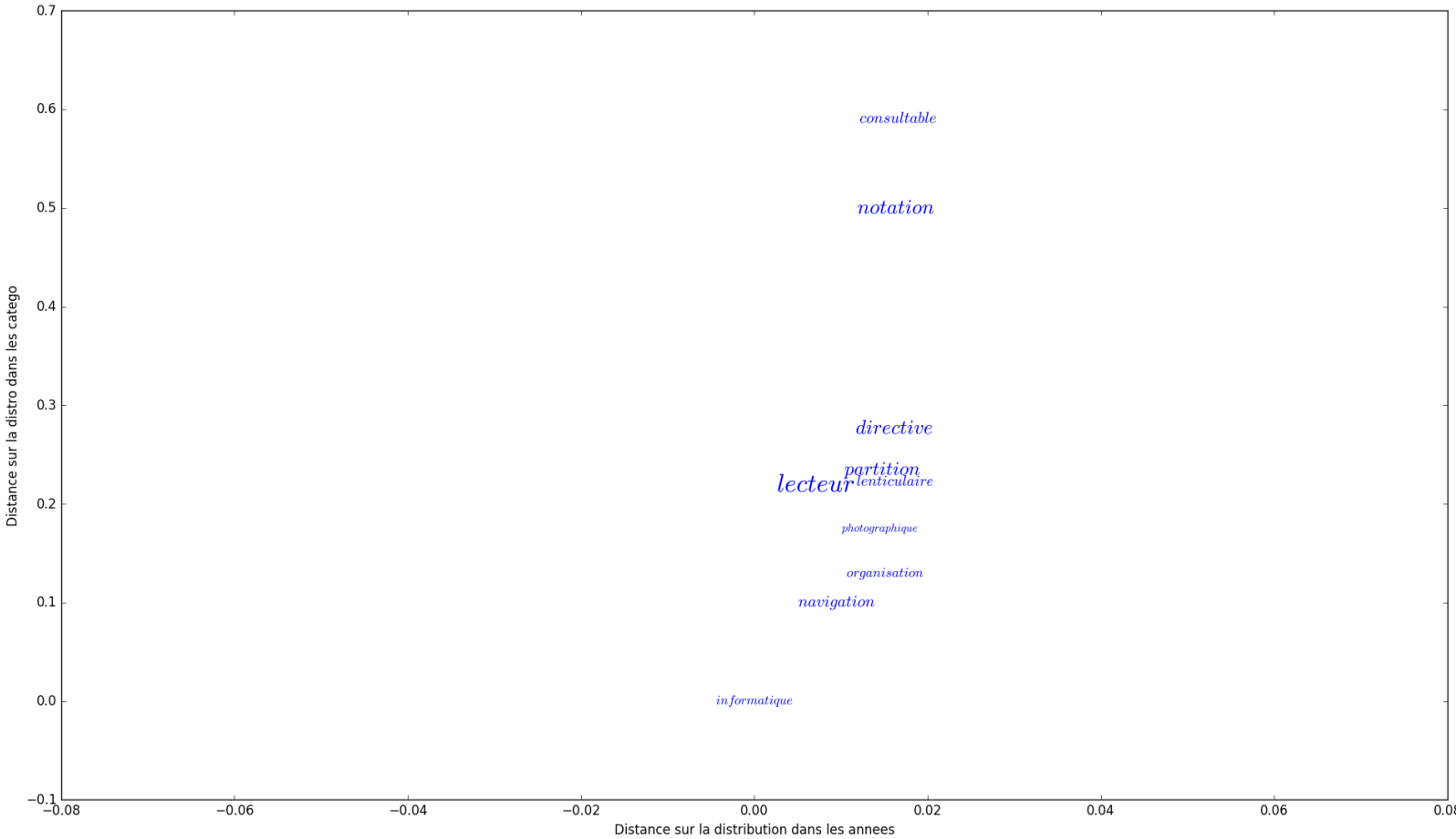
Créer une distance :

- Kullback-Leibler/euclidienne sur les deux histogrammes ;
- prendre le max.

Calcul de la distance à la volée sur tout le voc : 2s d'exécution

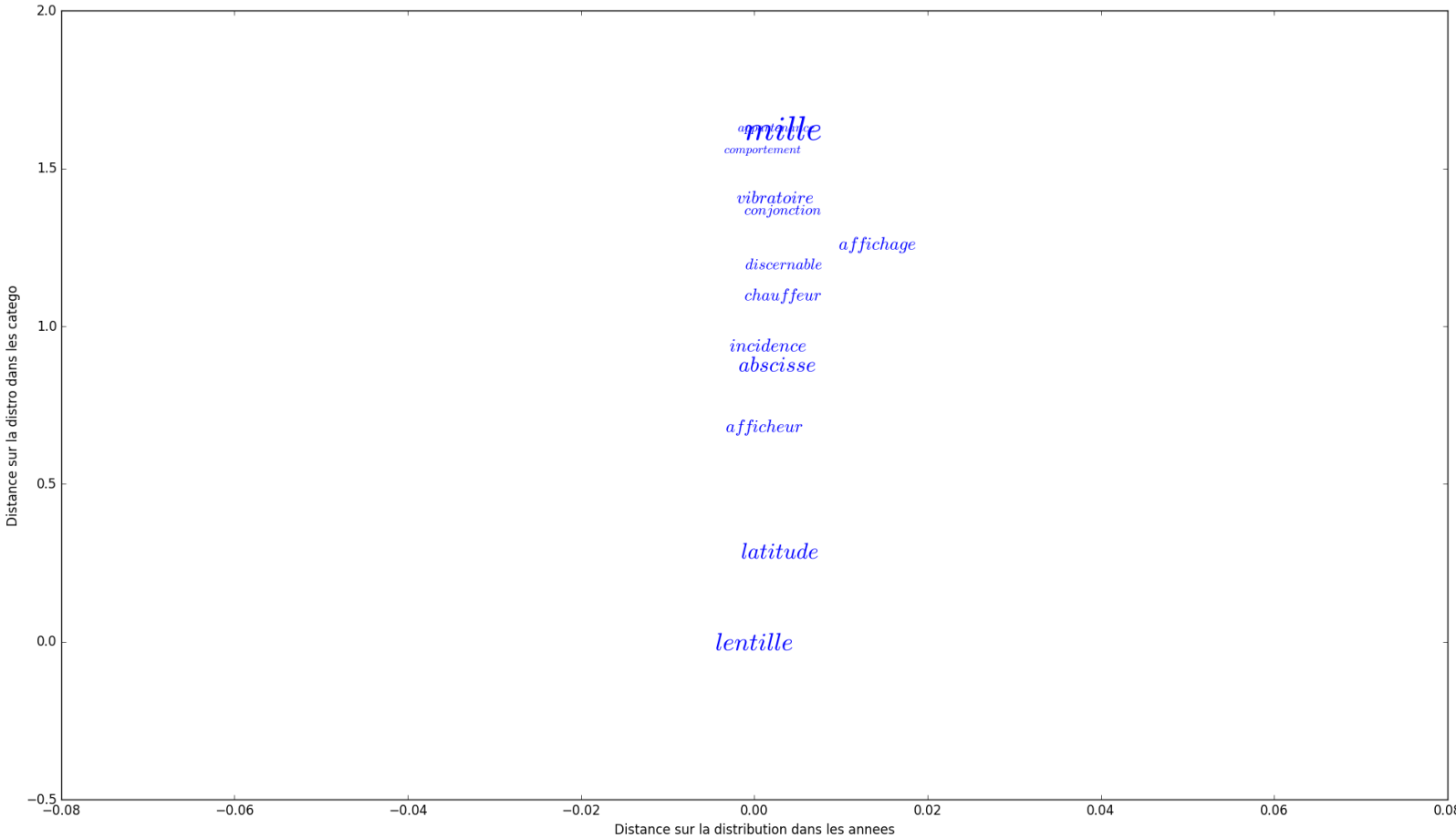
Résultat 1/2

```
python plotDistance.py informatique
```



Résultat 2/2

```
python plotDistance.py lentille
```



Explorer les co-évolutions 2/2

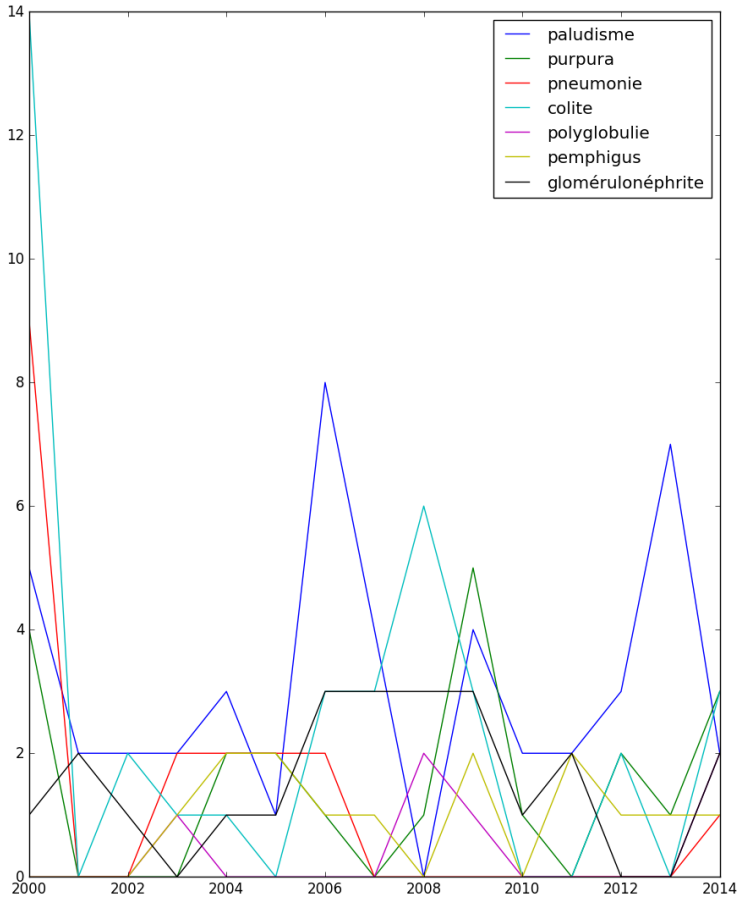
Distances sémantiques/contextuelles

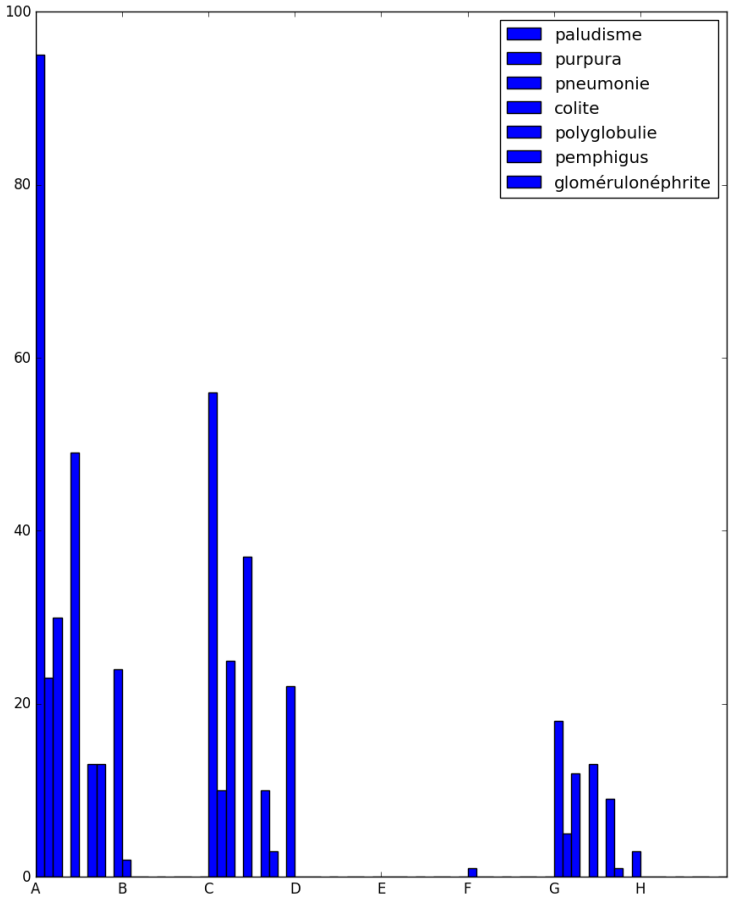
Utiliser la proximité selon un pré-calcul de *word embeddings* sur le corpus via Gensym.

Intuition : deux mots proches sont susceptibles d'évoluer similairement.

Résultat 1/3

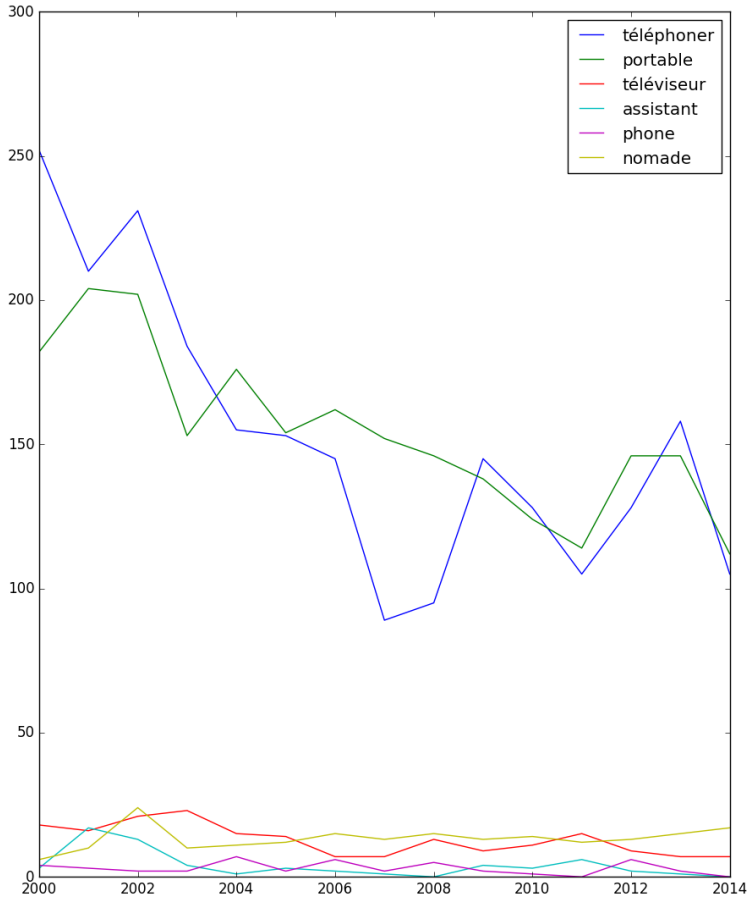
```
python3 neighbours.py sida
```

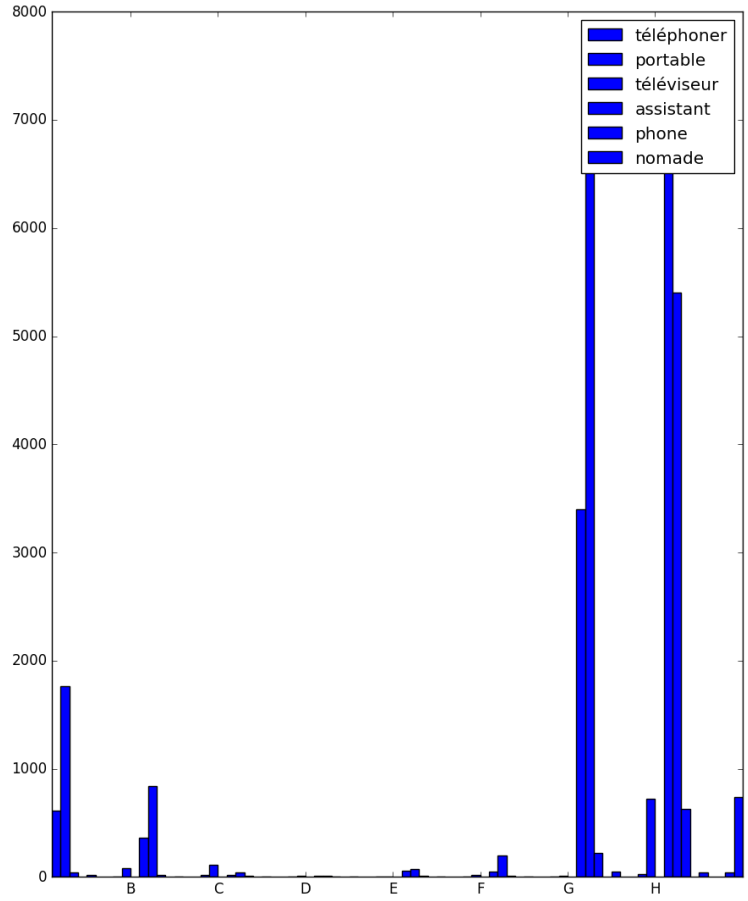




Résultat 2/3

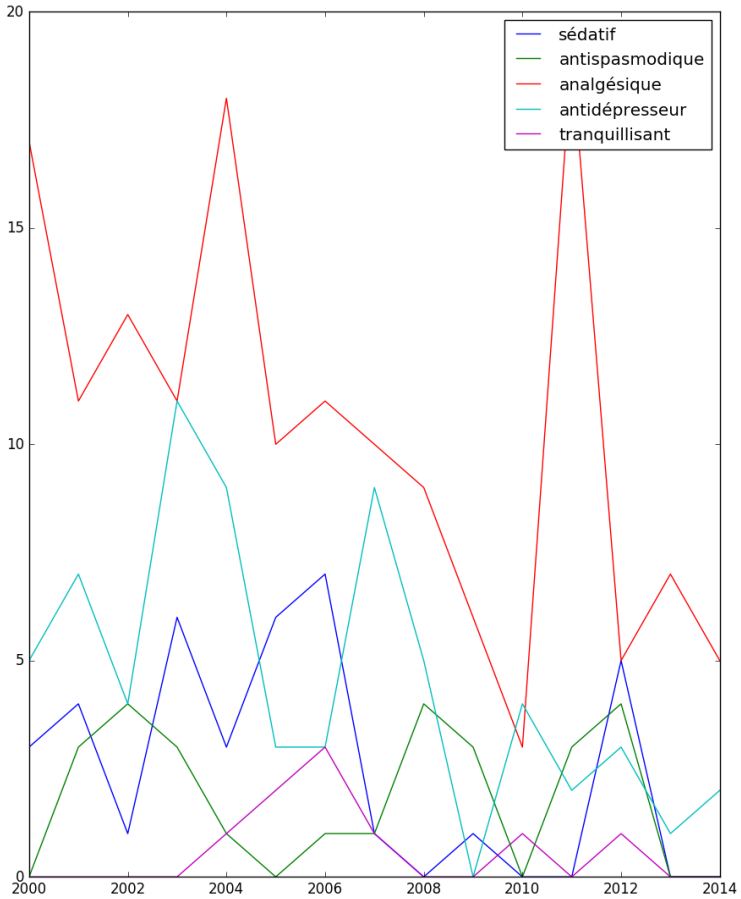
```
python3 neighbours.py smartphone
```

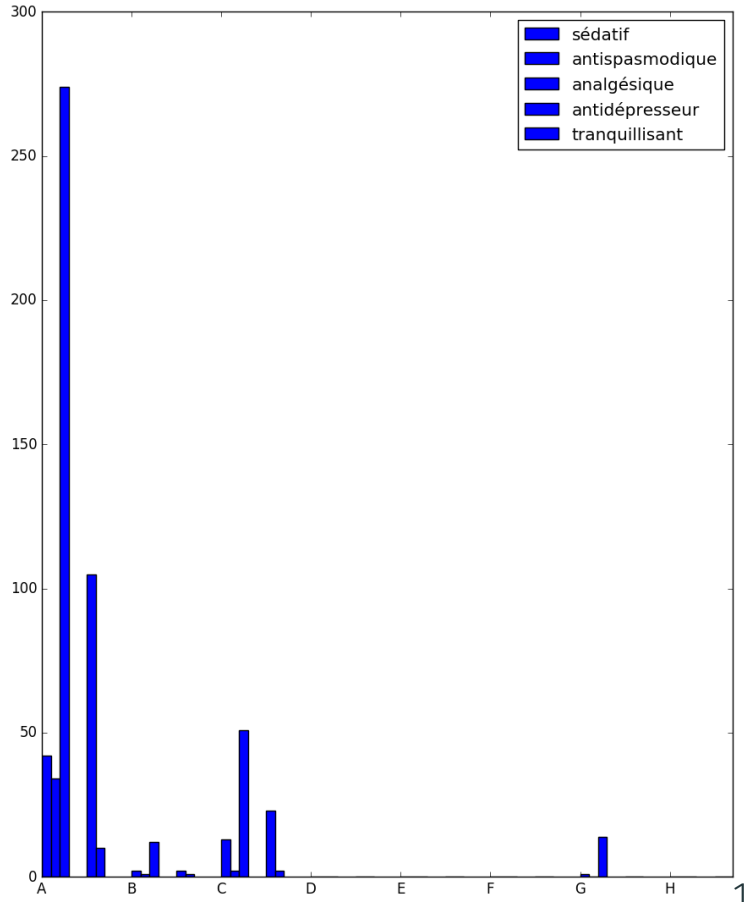




Résultat 3/3

```
python3 neighbours.py anxiolytique
```





Prendre de la hauteur 1/2 : caractériser le vocabulaire globalement

TF IDF

TF IDF adapté pour travailler sur les catégories/années

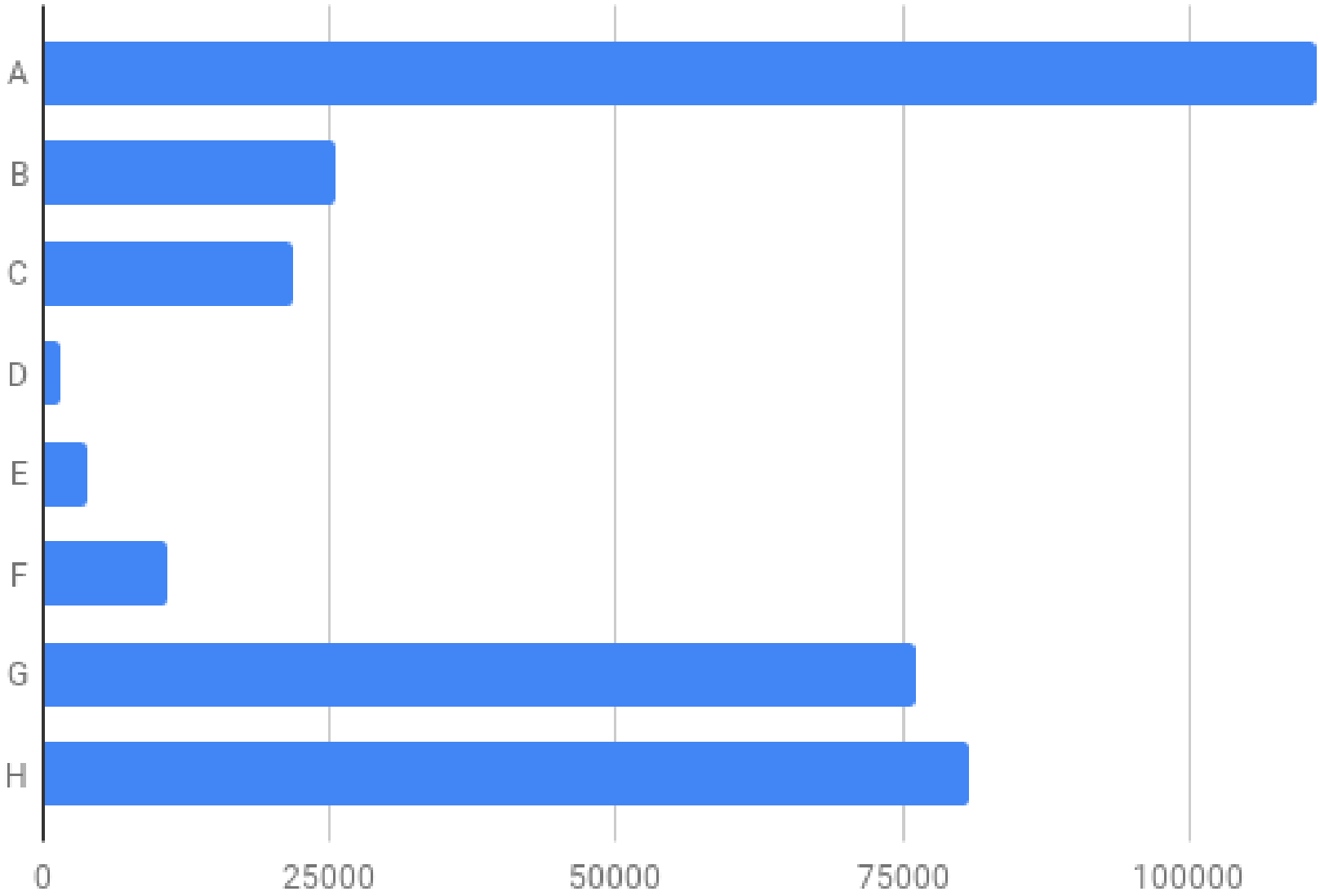
- TF I Categorie F
- TF I Annee F

TF IDF

```
 java -jar Hackatal-1.0-SNAPSHOT-jar-with-dependencies.jar karaoké
11:36:11.287 [main] INFO fr.labo.hackatal.Main - term: karaoké
11:36:11.521 [main] INFO fr.labo.hackatal.Main - Classement des années
2004 5.653478907336214E-4
2003 3.984328899908522E-4
2009 3.0319091725351446E-4
2005 2.718328067965136E-4
2002 1.3092745664022284E-4
2001 0.0
2012 0.0
2011 0.0
2010 0.0
2008 0.0
2007 0.0
2006 0.0
2015 0.0
2014 0.0
2013 0.0
11:36:11.590 [main] INFO fr.labo.hackatal.Main - Classement des domaines
G 7.635681634432142E-4
H 2.0608424386708314E-4
A 0.0
B 0.0
C 0.0
D 0.0
E 0.0
```

Prendre de la hauteur 2/2 : caractériser les clusters

Caractériser les clusters : A, G, H



Caractériser les clusters : A, G, H

Détecter les mots d'un clusters qui sont :

- **Représentatifs** du cluster ;
- **Saillants**, spécifiques au cluster.

$\rightarrow$ Feature F-mesure de Lamirel et al.

Caractériser les clusters : A, G, H

Détecter les mots d'un clusters qui sont :

- **Représentatifs** du cluster ;
- **Saillants**, spécifiques au cluster.

→ Feature F-mesure de Lamirel et al.

Découper le corpus en 5 périodes de 3 ans : est-ce que ce vocabulaire évolue ? → Approche **diachronique**.

Caractériser les clusters : A, G, H

```
   python comparaisonDecile.py specificites20012003A.dfsl
   specificites20042006A.dfsl

--------Mots stables très représentatifs----------
trifluorométhoxy
alkyléthersulfates
exponentiation
imides
intervertébral
[...]

--------Mots qui burst----------
altitudes
chondroïtine
rachis
arthrose
ylméthyl
extemporané
dopées
renseigne
moignon
agonistes
hydroxystéarate
urinaires
[...]
```

Conclusion

Conclusion

- Un gros jeu de données = de gros pré-traitements ;
- Des distances simples $\rightarrow$ dégager des tendances ;
- Les embeddings : utiles pour guider l'exploration ;
- Vision gros grain des clusters et années : nécessitent outils de visualisation fins mais des mots qui font le socle, d'autres qui burst/disparaissent.

Questions?

Backup slides

Ceci est un faux slide de BkacUp parce qu'on a bien entendu pas eu le temps d'en faire.

References i