Équipe “Broken pipe” : analyse lexicale des communiqués de presse

Sosie chauve de Yoann Dupont

Maître de conférences à Université Paris 3, LATTICE

Simon Guillot

Doctorant à Le Mans Université, LIUM

Nicolas Dugué

Maître de Conférences à Le Mans Université, LIUM

Europe, gauche, écologie, féminisme
-> Pas le même sens pour le RN que pour LFI

Comment tenter d'objectiver l'analyse lexicale ?

-> Plongements lexicaux

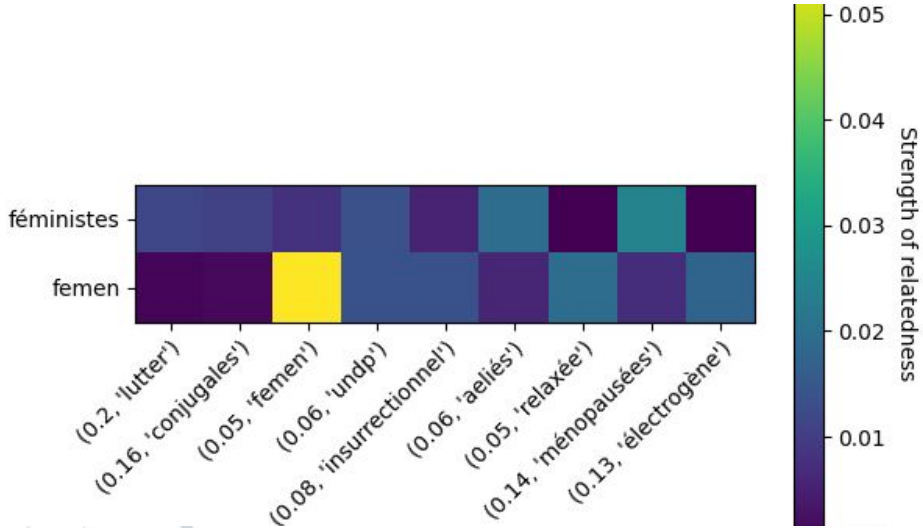
-> Analyses lexicométriques

Problème : petits corpus pour une telle tâche

Approche par transfert de plongements interprétables

> pip install sinr

Apprentissage de plongements SINr sur corpus média Leipzig



> Des bizarreries, modèle améliorable

Approche par transfert de plongements interprétables

> pip install sinr

Apprentissage de plongements SINr sur corpus média Leipzig

Transfert en utilisant :

- Dimensions obtenues sur Leipzig corpus
- Cooccurrences obtenues sur corpus de communiqués de presse

Les 10 co-occurrences les plus freq de loi sur les CDP LFI :

```
[65]: sorted(dico.items(), key=lambda x:x[1], reverse=True)[:10]
```

```
[65]: [('finances', 19),
       ('sécurité', 18),
       ('visant', 15),
       ('«', 11),
       ('plusieurs', 10),
       ('a', 9),
       ('programmation', 9),
       ('respect', 9),
       ('globale', 9),
       ('constitutionnelle', 8)]
```

Peut-on en fait quelque chose ? C'est TRÈS TRÈS PEU de stats pour du plongement :(

LIUM
Laboratoire d'Informatique
Le Mans Université

Lattice

[66]:
```
import numpy as np
vector = model.transfert_hackatal("loi", dico)
model.most_similar_vector("loi",vector.reshape(1, -1))
```

[66]:
```
{'object ': 'loi',
 'neighbors ': [('intérieure', 0.37),
  ('rectificative', 0.37),
  ('publiques', 0.37),
  ('rendues', 0.35),
  ('terrorisme', 0.32),
  ('lutter', 0.32),
  ('périmètre', 0.31),
  ('garantir', 0.31),
  ('cybercriminalité', 0.31),
  ('criminalité', 0.31),
  ('matière', 0.31),
  ('prévention', 0.31),
  ('glissières', 0.31),
  ('assurer', 0.31),
  ('renforcée', 0.31),
  ('efficacement', 0.31),
  ('lutte', 0.3),
  ('délinquance', 0.3),
  ('insurge', 0.3),
  ('failles', 0.3),
  ('protestent', 0.29),
  ('protester', 0.29),
  ('assainir', 0.29),
  ('fléau', 0.29)]}
```

Crise - CDP FI

```
model.most_similar_vector(mot,vector.reshape(1, -1))
```

```
{'object ': 'crise',
 'neighbors ': [('cordon', 0.46),
  ('anses', 0.33),
  ('invs', 0.31),
  ('sociale', 0.27),
  ('environnementale', 0.26),
  ('crise', 0.26),
  ('cohésion', 0.25),
  ('inclusion', 0.24),
  ('envahissants', 0.24),
  ('burn', 0.24),
  ('sanitaire', 0.24),
  ('mixité', 0.23),
  ('conjoncture', 0.23),
  ('travailleuse', 0.23),
```

Crise - CDP RN

```
model.most_similar_vector(mot,vector.reshape(1, -1))
```

```
{'object ': 'crise',
 'neighbors ': [('migratoires', 0.4),
  ('crise', 0.38),
  ('cordon', 0.36),
  ('conjoncture', 0.34),
  ('marasme', 0.31),
  ('flux', 0.29),
  ('traverse', 0.29),
  ('inclusion', 0.29),
  ('secoue', 0.27),
  ('sociale', 0.27),
  ('feedly', 0.27),
  ('migratoire', 0.27),
  ('agrégateur', 0.26),
  ('anses', 0.26),
```

Répression- CDP FI

```
{'object ': 'répression',
 'neighbors ': [('lutte', 0.67),
  ('terrorisme', 0.65),
  ('protestent', 0.65),
  ('protester', 0.65),
  ('corruption', 0.63),
  ('luttant', 0.62),
  ('luttons', 0.62),
  ('criminalité', 0.61),
  ('fléau', 0.61),
  ('cybercriminalité', 0.61),
  ('insurge', 0.61),
  ('délinquance', 0.61),
  ('discriminations', 0.6),
  ('luttent', 0.6),
  ('apartheid', 0.6),
```

Répression - CDP RN

Finance - insécurité - fermeté

```
: {'object ': 'répression',
   'neighbors ': [('brdp', 0.36),
    ('fraudes', 0.34),
    ('dgccrf', 0.34),
    ('sanglante', 0.33),
    ('délinquance', 0.3),
    ('massives', 0.29),
    ('sévère', 0.27),
    ('insécurité', 0.25),
    ('fermeté', 0.24),
    ('implacable', 0.24),
    ('criminalité', 0.24),
    ('pacifiques', 0.24),
    ('révolte', 0.23),
    ('malnutrition', 0.23),
    ('brutale', 0.23),
```

Quelle distance entre les concepts dans les communiqués de presse FI-RN ?

Sur une liste de cibles arbitraires :

```
[('crise', 0.28362747864463 58), ('liberté', 0.34413778755109 58), ('loi', 0.4488310
7531200606), ('européenne', 0.5564079436494442), ('macron', 0.5912512540333839),
('répression', 0.7148312876054919), ('emploi', 0.7864057801840947), ('gauche', 0.9
162067185949125)]
```

LIUM
Laboratoire d'Informatique
Le Mans Université

Lattice

Poster de Simon sur les plongements interprétables

De l'interprétabilité des dimensions à l'interprétabilité du vecteur : parcimonie et stabilité

Guillot Simon, Prouteau Thibault, Dugué Nicolas

Venez nombreux·ses !

Les termes spécifiques communs à au moins deux partis

- Corpus presse ⇒ partition par parti politique
- Calcul spécificité mots : loi hypergéométrique (Lafont)
- Quels mots sont spécifiques à au moins deux partis ?
- Avec les pôles communs
  - Retour au texte pour voir traitement spécifique
  - Pas le temps pour des calculs de cooccurrence

Termes communs LI - FN

- Spécificité > 1 pour les deux (68 mots)
- Pas de tri après
- Vérifications à faire pour chaque terme

france	54.3067	92.6074
toute	1.0353	3.1248
commission	7.2888	25.9327
Communiqué	26.0667	13.0687
presse	23.5237	1.7265
député	3.9948	11.1042
état	2.253	18.5916
devant	1.1811	2.5767

députés	16.2088	1.1261
macron	14.2857	14.0794
aucune	2.4452	2.2488
elles	1.6911	2.2505
parlement	2.6994	19.2598
jean	1.2623	29.6186
police	1.7114	1.923

Nombre de termes spécifiques avec spécificité > 1

FI - LR ⇒ 183 FI - PS ⇒ 490 FI - RN ⇒ 68	LR - PS ⇒ 395 LR - RN ⇒ 44