Premiers résultats

Claire He

Réunion du 10/09

Résumé

- Scrapping automatisé : exécuter le fichier **main.py** du dossier *scrapping* dans le terminal de commande python.
- Stat desc, nettoyage complémentaire (voir notebook html)
- Réimplémentation de l'approche *bag of words* pour la LDA avec *scikit-learn*, comparaison avec techniques NMF
- Pistes bibliographiques

Scrapping automatisé

Fonctionnement sur terminal/console de commandes python3

Télécharger le dossier *scrapping*. Après exécution en suivant les commandes suivantes, tous les transcripts de la FOMC à partir de la date renseignée par l'utilisateur dans le script jusqu'aux derniers publiés. Détails dans le fichier README.md.

- Installer python3
- Ouvrir le terminal ou la console
  - macOS : > dossier Utilitaires > *Terminal*
  - Windows : > *Invite de commandes*
  - Linux : > *Terminal* ou Ctrl+Alt+T
- Aller dans le dossier du package *scrapping* (voir détail ci contre)
- Taper *python3 main.py*

```
(base) MBP-de-H2JW:scrapping h2jw$ python3 main.py
100%|████████████████████████████████████| 158/158 [00:04<00:00, 36.50it/s]
100%|████████████████████████████████████| 142/142 [00:03<00:00, 46.86it/s]
100%|████████████████████████████████████| 229/229 [00:07<00:00, 30.63it/s]
100%|████████████████████████████████████| 256/256 [00:07<00:00, 34.48it/s]
100%|████████████████████████████████████| 262/262 [00:07<00:00, 34.47it/s]
100%|████████████████████████████████████| 184/184 [00:03<00:00, 46.97it/s]
100%|████████████████████████████████████| 216/216 [00:04<00:00, 50.52it/s]
100%|████████████████████████████████████| 175/175 [00:03<00:00, 54.90it/s]
100%|████████████████████████████████████| 299/299 [00:07<00:00, 42.00it/s]
100%|████████████████████████████████████| 191/191 [00:04<00:00, 39.59it/s]
```

Pour macOS et Linux (UNIX systèmes) : commandes en bash

- Pour dérouler le dossier courant : *ls*
- Pour accèder à un dossier : *cd*

```
(base) MBP-de-H2JW:~ h2jw$ ls
Applications            Music                   TP6_Modelisation.ipynb
Desktop                 Pictures                Untitled.ipynb
Documents               Projet_ST.Rmd           Zotero
Downloads               Projet_ST.nb.html       nltk_data
Dropbox                 Public                  opt
Library                 TP03_geopandas.ipynb    scikit_learn_data
Movies                  TP04_webscraping.ipynb  seaborn-data
(base) MBP-de-H2JW:~ h2jw$ cd Documents/GitHub/NLP-FOMC/scrapping
```

Dossier parent

Dossier parent

Accès au package : *cd + chemin du dossier scrapping*

Pour Windows :

- Pour dérouler le dossier courant : *pwd* ou *cwd*
- Pour accéder à un dossier : *cd*

LDA : pistes et commentaires

Perspective : topic modelling et influence des chairs/chairman/audience ?
Ouverture : autre technique de topic modelling NMF en annexe dans les résultats

- LDA sur tout le corpus en variant le nombre de topics
  - Approfondissement possible : nombre de topics optimal via clustering T-SNE/K-means après l'approche *bag of words* avec TF-idf

- LDA par chair et sur les personnes en tandem entre deux chairs
  - Observation via pyLDAvis d'un possible impact (apparition ou disparition d'un topic conséquent)
  - Critique : corpus trop petit, donc résultats non statistiquement significatifs, à voir sur le point suivant

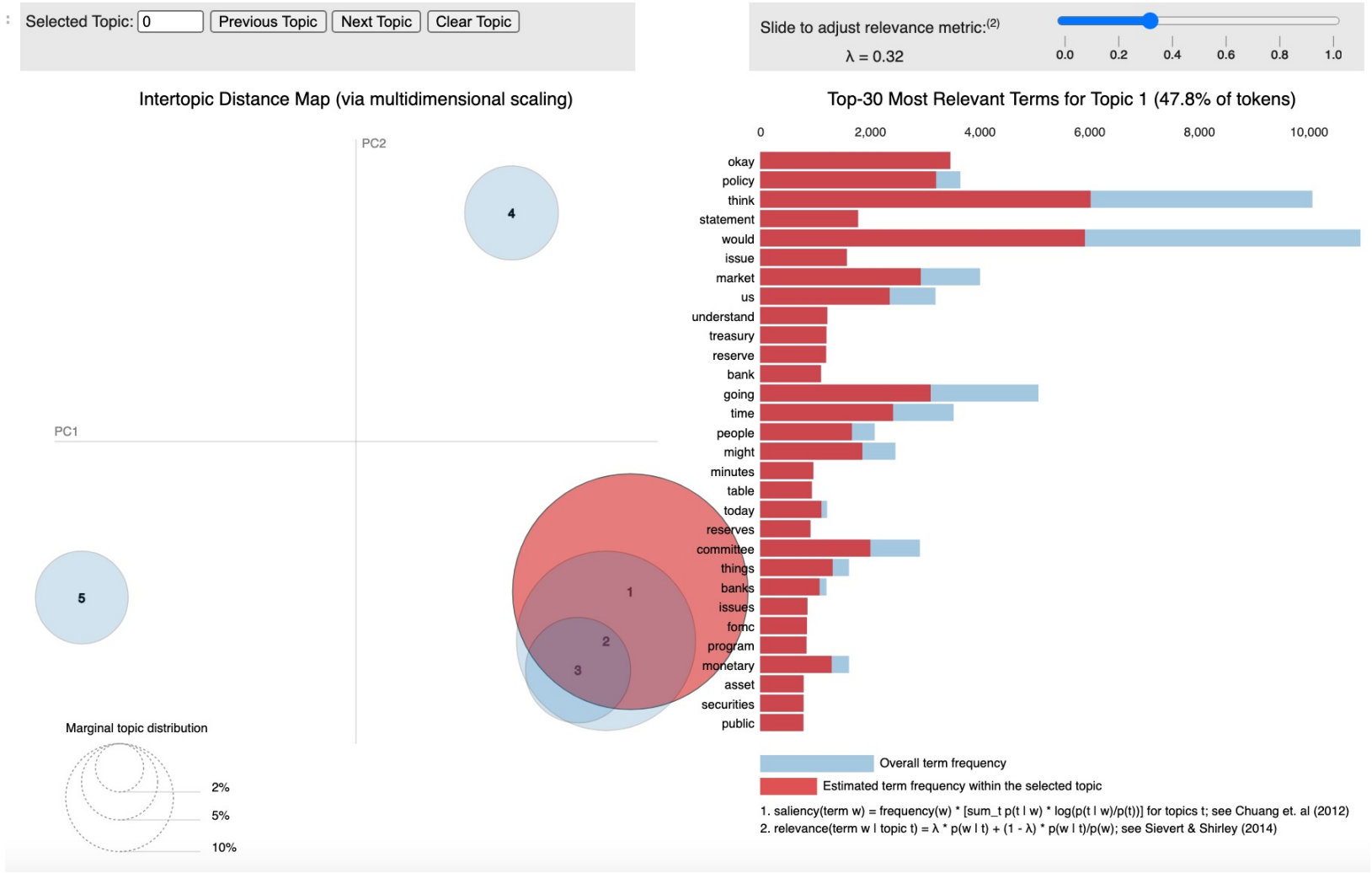
- LDA sur tout le corpus, observation de la fréquence des topics par chair
  - Ouverture : observation des fréquences pour les personnes entre deux chairs des topics obtenus sur « tout le corpus »

LDA : pistes et commentaires

LDA sur un extrait de corpus (5000 lignes)

- 3 zones dédiées à la terminologie macro
- 2 zones correspondant à la structure de la réunion (tours de parole etc.)

Problème à explorer : Choix de la bonne métrique pour évaluer la pertinence d'un topic modelling



LDA : pistes et commentaires

De la **chair Greenspan :** évolution au passage à **Bernanke** ?
LDA sur les individus à cheval sur les deux mandats.

Greenspan

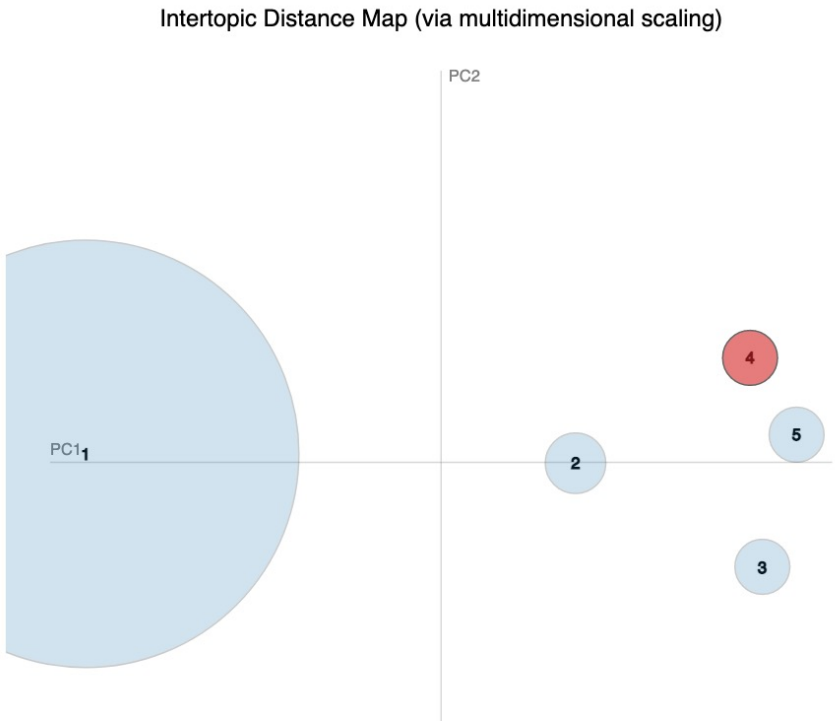
1 : rate growth percent inflation year policy us panel prices forecast

2 : press commitee saving recall swap assets figure budget break raise banking

3: signal recommendation compensation uncertainties reports wage bluebook power

4: results parts series probability momentum budget orders

5: previous maintain lead record reserves product export consequence



Bernanke

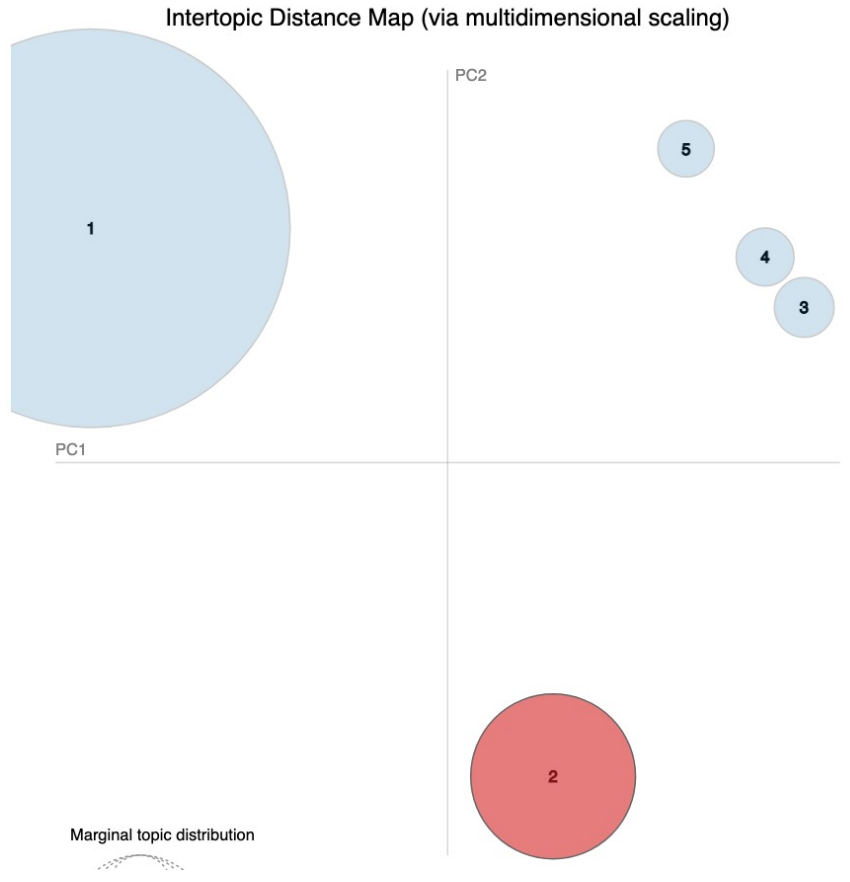
1 : rate inflation market funds federal reserve securities financial risk treasury credit

2 : central projections percent unemployment gdp growth forecast panel assessments greenbook

3: production management materials reducing estimates extent positive side chart

4: liftoff premiums overnight moreover memo implication probability deleveraging commodity

5: remittances sheet board repo judge look rather facility structural stress scenario weaker



LDA : pistes et commentaires

De la **chair Greenspan :** évolution au passage à **Bernanke** ?
LDA sur les individus à cheval sur les deux mandats.
Idem pour Yellen/Bernanke

Greenspan

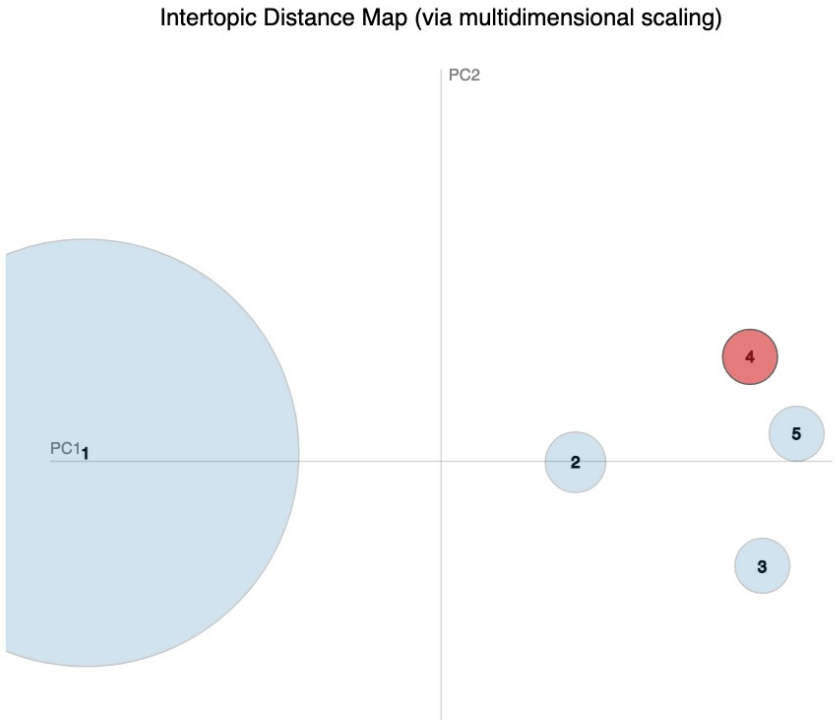
1 : rate growth percent inflation year policy us panel prices forecast

2 : press commitee saving recall swap assets figure budget break raise banking

3: signal recommendation compensation uncertainties reports wage bluebook power

4: results parts series probability momentum budget orders

5: previous maintain lead record reserves product export consequence



Bernanke

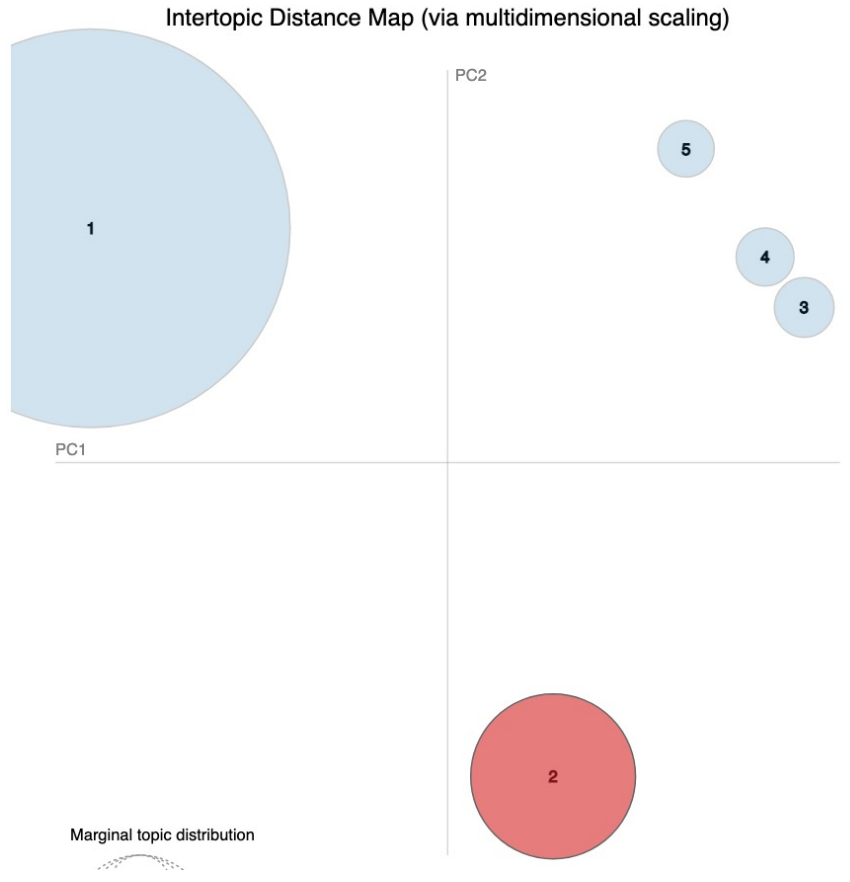
1 : rate inflation market funds federal reserve securities financial risk treasury credit

2 : central projections percent unemployment gdp growth forecast panel assessments greenbook

3: production management materials reducing estimates extent positive side chart

4: liftoff premiums overnight moreover memo implication probability deleveraging commodity

5: remittances sheet board repo judge look rather facility structural stress scenario weaker



LDA : pistes et commentaires

LDA sur tout le corpus. Observation des topics dominant selon les chair :

Topic 1: Inflation et conjoncture économique

ex. mots clés : *growth, percent, inflation, year, rate, prices, forecast, unemployment, labor, economy*

Topic 2: Aspects financiers et régulation

ex mots clés : *president, market, governor, purchases, chairman, billion, asset, meeting, december, bank*

Topic 3: Anticipation de la réponse d'une politique monétaire (?)

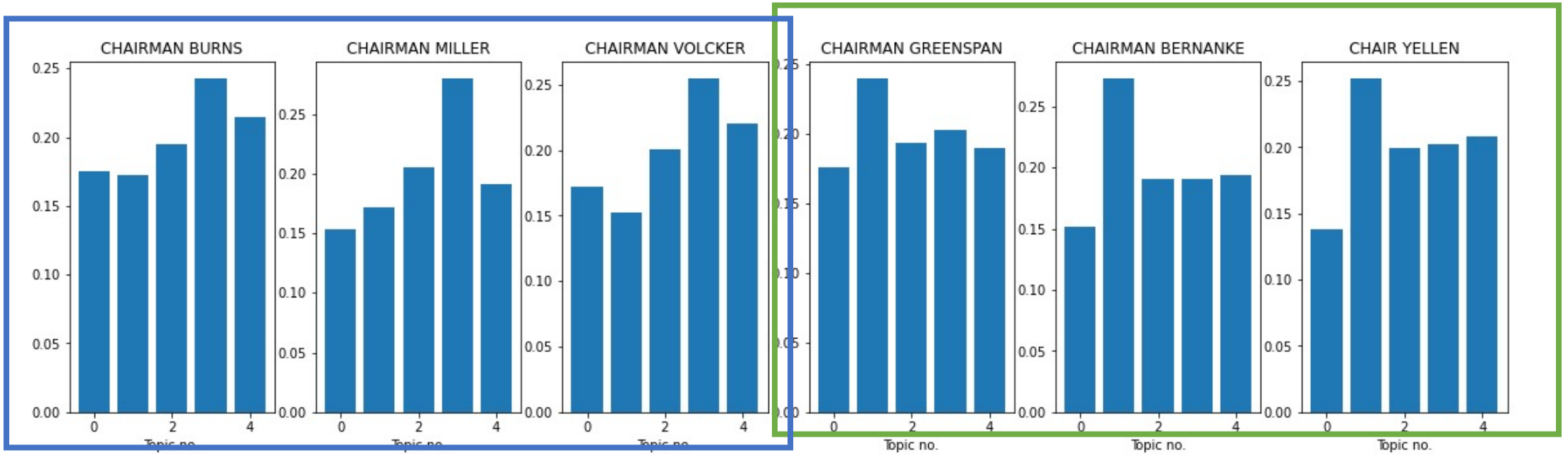
Ex mots clés : *policy, inflation, rates, monetary, right time, economy, expectations, going*

Topic 4: Opposition des points de vue/apport d'idée

Ex mots clés : *alternative, percent, statement, committee, rate, funds, language, point, want*

Topic 5: Marchés financiers et consommateurs

Ex mots clés : *market, going, financial, people, lot, little, credit, number, markets*



Pistes et bibliographie :

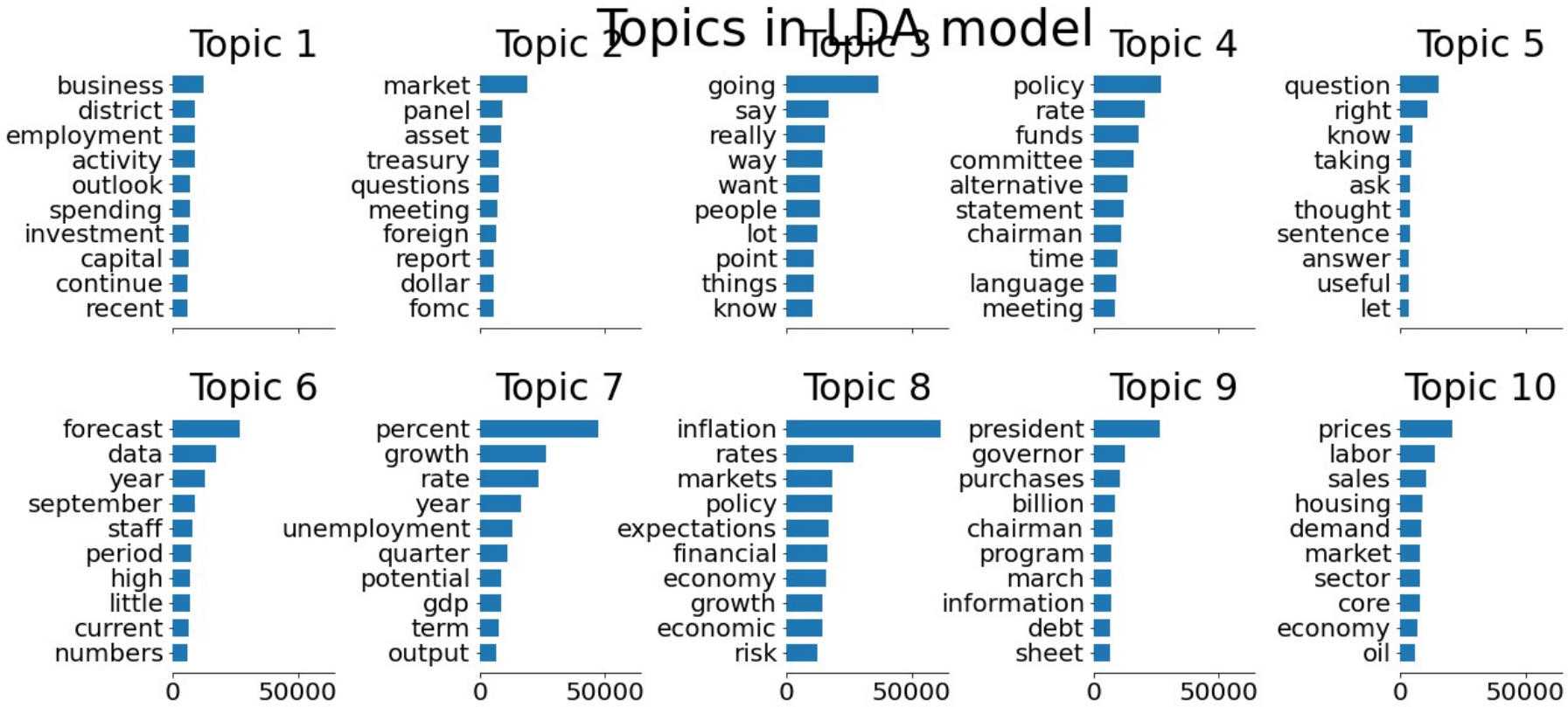
- Perspectives complémentaires sur l'article sur la Révolution Française : avec détail de la modélisation de la novelty, transience and resonance et LDA https://www.pnas.org/content/pnas/suppl/2018/04/16/1717729115.DCSupplemental/pnas.1717729115.sapp.pdf
- LDAvis : https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf (intègre du clustering pour paramètrer le modèle)
- LDA avec online learning (pour le paramétrage) https://papers.nips.cc/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf (pour l'implémentation) https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html
- ouverture sur autre modèle de topic modelling non supervisé : https://medium.com/voice-tech-podcast/topic-modelling-using-nmf-2f510d962b6e
- Ouverture: clustering pré-LDA pour définir le nombre de topics : https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5
- LDA2VEC : modèle mixte : https://github.com/meereeum/lda2vec-tf implémentation https://arxiv.org/abs/1605.02019 papier

Résultats annexes

Divers tests exploratoires

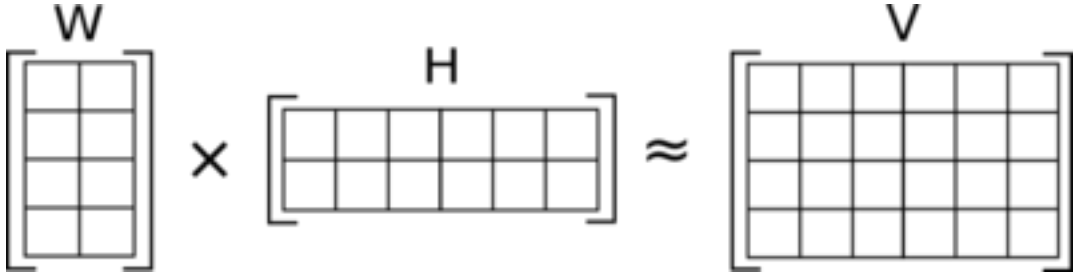
Résultats : LDA à 10 topics

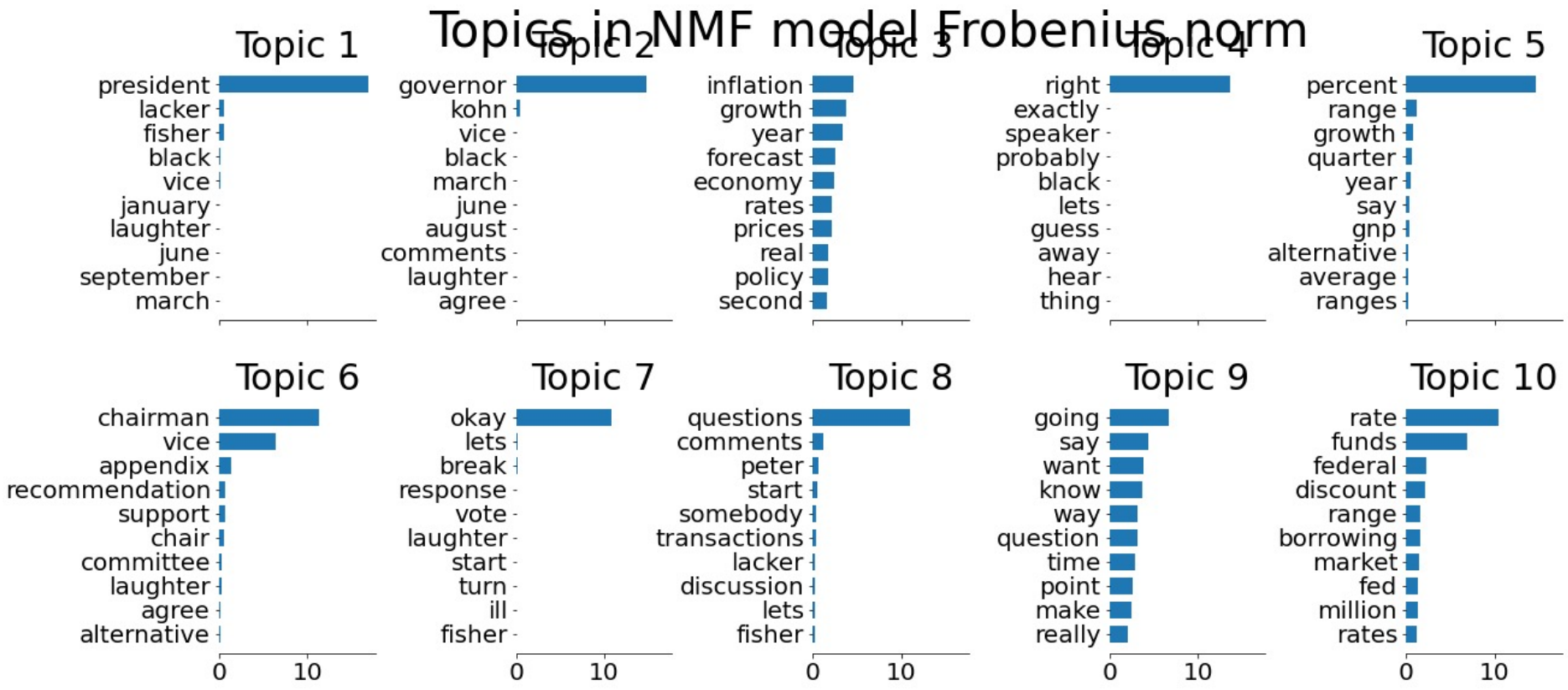
Réimplémentation LDA *bag of words*



Résultats : NMF

Non-Negative Matrix Factorization (NMF) :

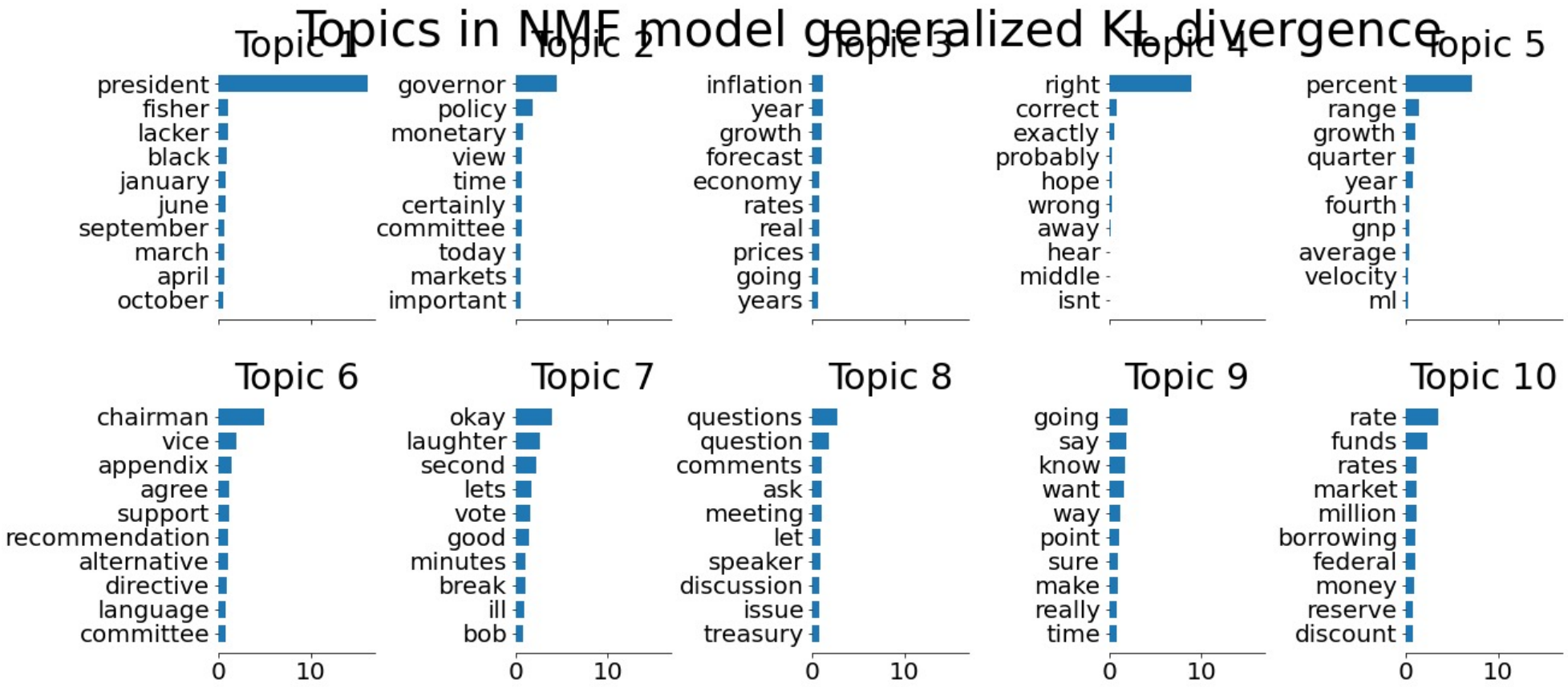




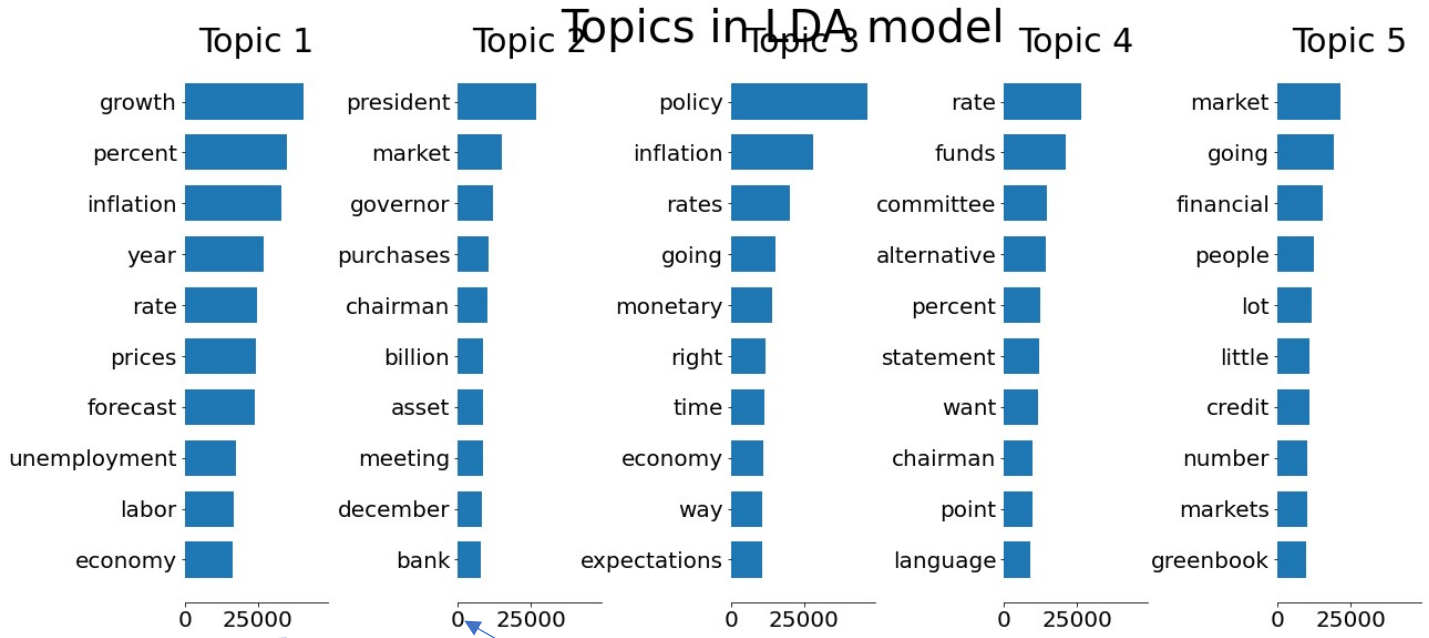
Résultats : NMF

$$\text{kl\_div}(x, y) = \begin{cases} x \log(x/y) - x + y & x > 0, y > 0 \\ y & x = 0, y \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

Non-Negative Matrix Factorization (NMF) :



LDA à 5 topics exécutée sur l'ensemble du corpus de statements



Comparaison au travail d'Etienne

Le dataset a été mis à jour sur l'année 2015.
Je n'ai pas réussi à exploiter les codes d'Etienne sur la LDA (pas trouvé le code correspondant).

Résultats d'Etienne

Topic 0 : president market governor policy committee
Topic 1 : percent rate growth time range
Topic 2 : inflation policy market economy financial
Topic 3 : forecast prices gdp market unemployment data
Topic 4 : economy year time district people