La modélisation thématique non-supervisée (LDA) : Comment catégoriser des quantités massives de données textuelles?

Maxime Bérubé, PhD
Chercheur postdoctoral
Concordia University/University of Waterloo
maxime.berube@concordia.ca

GARDESO

Groupe sur l'Analyse, la Recherche et le Développement En Source Ouverte

Gardeso · Ressources · Activités · Nous joindre

Python – Gensim LDA topic modeling

Une fois les données nettoyées (dans le cas de tweets par exemple, retrait de caractères spéciaux, emojis, retours de chariot, tabulations, etc.), la modélisation thématique LDA à l'aide du module Gensim (Python) s'effectue par les 6 étapes suivantes:

1. Chargement des modules nécessaires (si ce n'est pas déjà fait)
2. Importation des données (si ce n'est pas déjà fait)
3. Pré-traitement des données (tokenization, lemmatization, stopwords)
4. Bag-of-words
5. Construction du modèle LDA
6. Visualier le modèle avec pyLDAvis

1. Chargement des modules

```
# S'assurer d'avoir le modèle spacy pour le text pre-proces
# Rouler les commandes suivantes dans un terminal au besoi
#pip install -U spacy
#py -m spacy download en
#py -m spacy download fr
#pip install pyLDAvis
#pip install -U textblob
#py -m textblob.download_corpora
```

POSTS

- Python – Gensim LDA topic modeling
- Python – Vérification de la cohérence de modèles LDA

Social Media Youtube CSV Recherche Topic modeling Sauvegarde WHOIS LDA Collecte Audience Stopwords Spacy Outils Gensim Facebook Alertes Tokenization Scraping Pandas Monitoring JSON OSINT Time series Comments Network Python Content Surveillance Numpy Site web Handbook pyLDAvis Twitter HTML Coherence score Lemmatization

Search ...

NOUVEAUTÉS

https://gardeso.com/python-gensim-lda-topic-modeling

Qu'est-ce que le *topic modeling*?

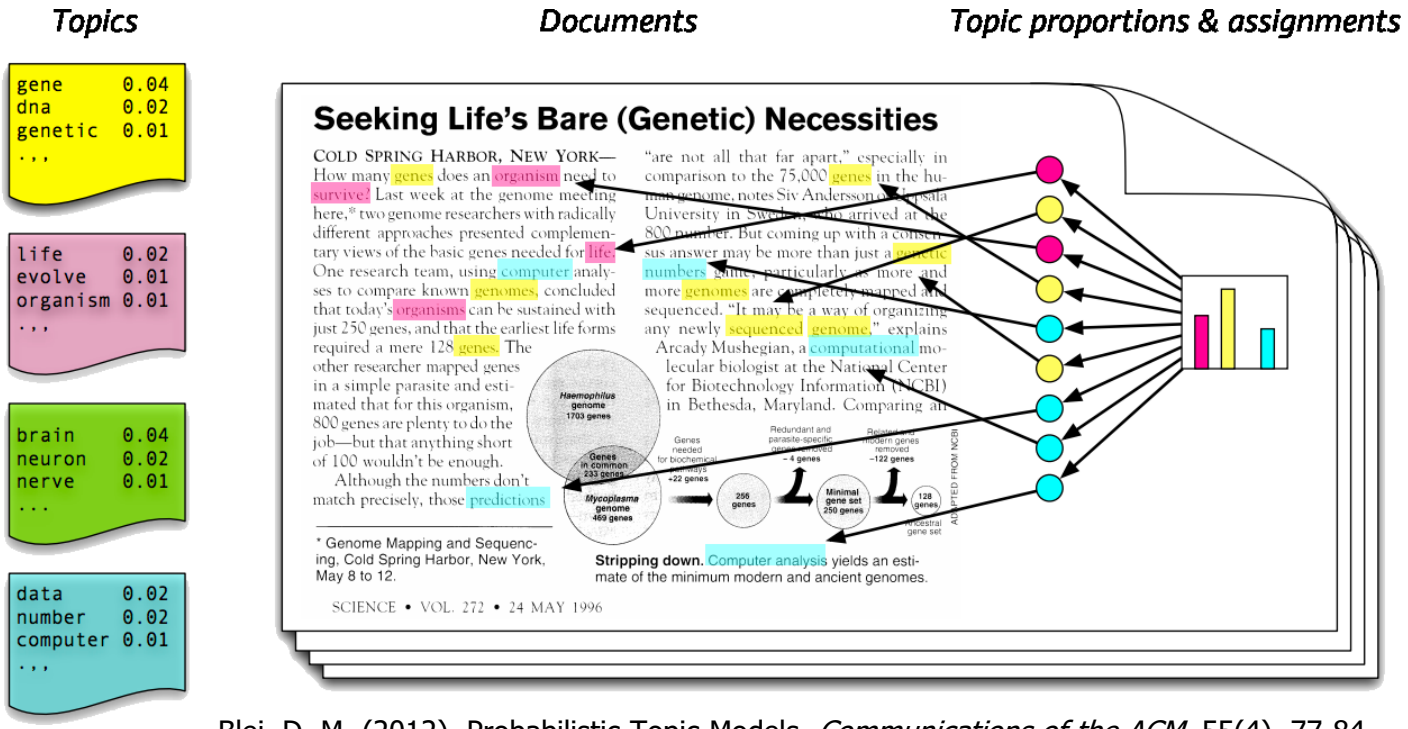
- Le *topic modeling* ou la modélisation thématique est un type d'analyse quantitative qui vise à mieux connaître le contenu d'une quantité massive de documents textes.

- Elle consiste en une technique de **traitement du language naturel** permettant d'**extraire des thèmes latents** d'un corpus de documents.

- Il s'agit d'un type d'analyse que l'on dit **non-supervisée**, c'est-à-dire que l'on n'a pas à connaître les thèmes au préalable. Ils sont dictés par l'analyse.

Les 4 composantes essentielles

1. Les documents;
2. Les thèmes (*topics*);
3. La distribution statistique des thèmes dans les documents;
4. Les mots associés à ces thèmes.

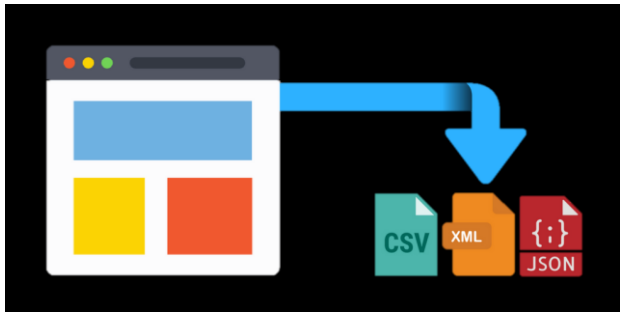


Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM,* 55(4), 77-84.

Les pré-requis d’une bonne analyse

1. La qualité des textes analysés
   - Le moins d'expressions familières, de fautes d’orthographe et de caractères spéciaux que possible.
2. La variété de thèmes proposés dans les textes.
   - Les documents doivent avoir une diversification interne minimale pour faciliter l'identification de thèmes différents.
3. Le choix de l'algorithme pour la modélisation.
   - LSA, HDP, NMF, LDA, etc.
4. Le nombre de thèmes identifiés par l'algorithme.
   - Il est possible de se fier à des indicateurs de cohérence et de perplexité afin de cibler le nombre minimal de thématiques avec la meilleure cohérence possible (p. ex. : c-value).
5. La préparation des données.
   - L'exclusion de *stopwords*, le *pre-processing*, etc. (Tokenization, lemmatization, bag of words)
6. Le bon réglage des paramètres de l'algorithme.
   - num_topics, random_state, passes, etc.

Le processus d'analyse

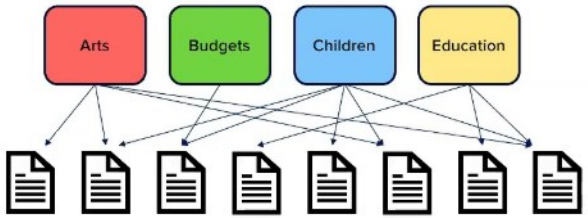


1. Collecte

python™

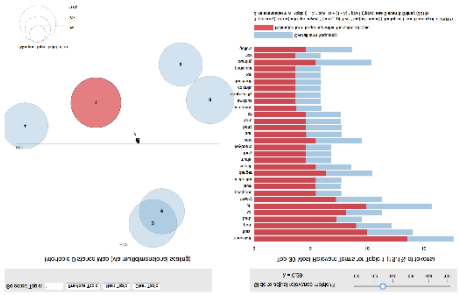
2. Importation et chargement des modules

3. Nettoyage et pré-traitement



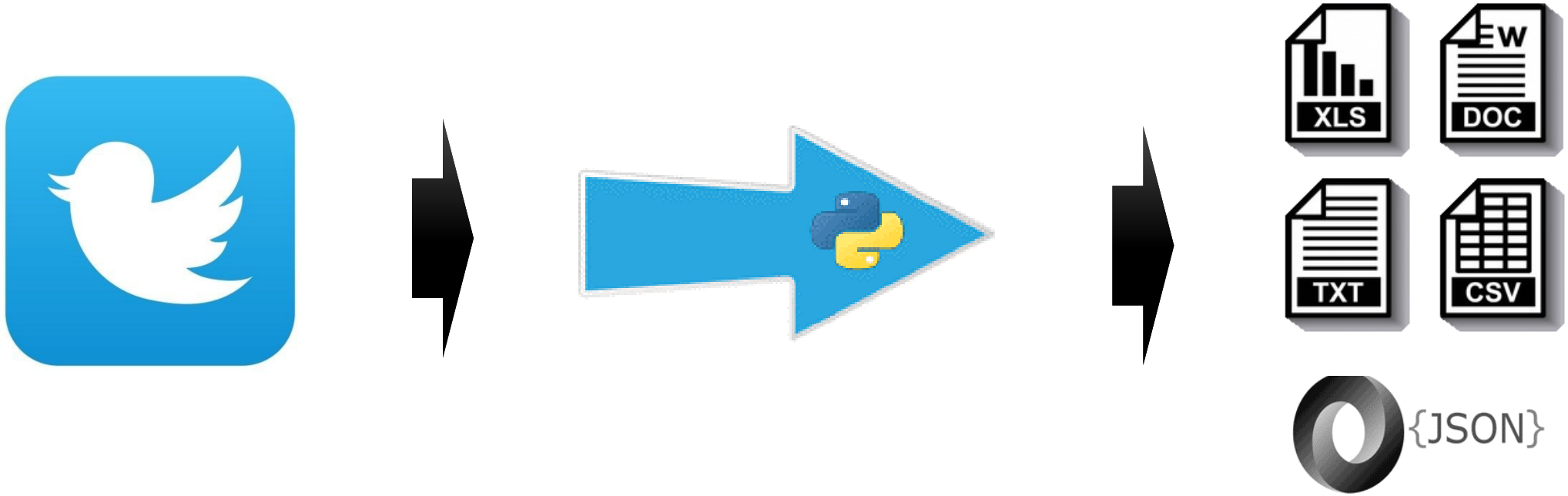
4. Construction du modèle LDA

5. Évaluation du modèle

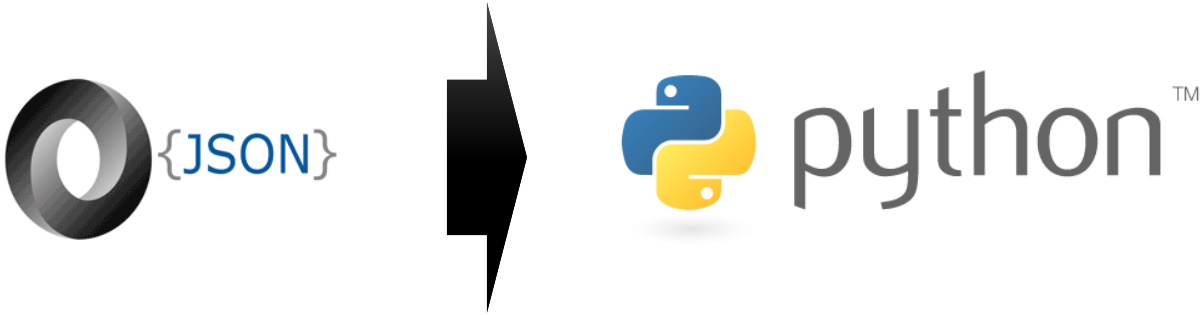


6. Visualisation

1. Collecte



2. Importation et chargement des modules



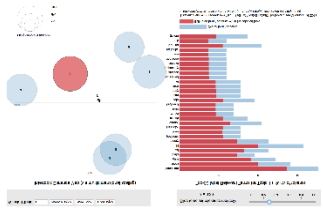
Construction et manipulation de tableau

Tokenization et lemmatization

Stopwords et bigrams

Modèle LDA

pyLDAvis



Visualisation

3. Nettoyage et pré-traitement

Nettoyage des données :

- Retrait des caractères spéciaux, ponctuations, en
  chariot, tabulations, hyperlinks/URLs, retweets, e
- Convertir le texte en minuscules.
- Convertir les abréviations en mots.
- Remplacer les numéros en texte.
- Réécrire proprement les mots allongés (p. ex.: « [illegible] »).
- Combiner les mots issus d'un usage d'anglais Americain VS British (p. ex.: color vs colour).

3. Nettoyage et pré-traitement

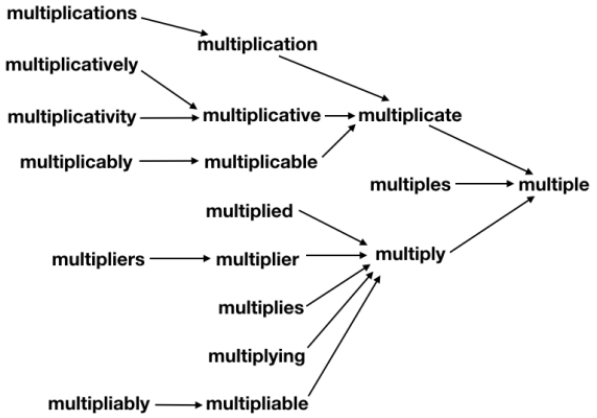
Tokenization :

- Séparer les phrases en suites de mots.

```
<sentence>
  <word>The</word>
  <word>quick</word>
  <word>brown</word>
  <word>fox</word>
  <word>jumps</word>
  <word>over</word>
  <word>the</word>
  <word>lazy</word>
  <word>dog</word>
</sentence>
```

Lemmatization :

- Associer les différentes formes de mots à leur racine lexicale, tout en considérant le contexte dans lequel ils se trouvent.
- Identifier noms, adjectifs, verbes, adverbes, etc.



Stopwords :

- Retirer les mots indésirables du corpus.
- Ajouter des mots à la liste au besoin.

• a	• of	• on
• I	• for	• with
• the	• at	• from
• in	• to	

3. Nettoyage et pré-traitement

Bag-of-words :

- On assigne un identifiant numérique à chacun des mots du corpus **(dictionnaire)** et on y indique sa fréquence par document (ou par ligne d'entrée (cellule) dans un fichier CSV) **(corpus)**.

Dictionnaire :

1, Statement
2, Incident
3, Arena
4, Concert
5, Explosion
6, Casualty
Etc.

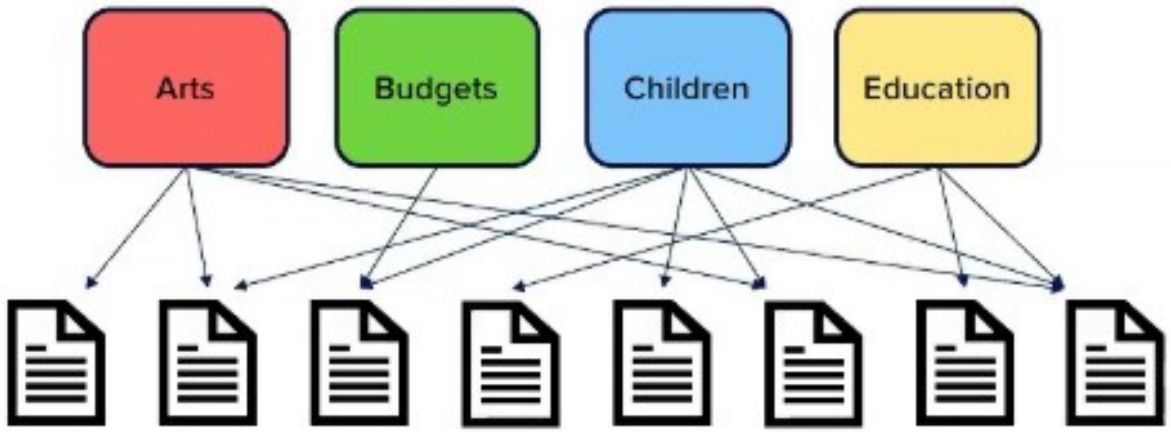
Corpus :

```
Out[11]: [(3, 1),
          (13, 1),
          (37, 1),
          (38, 1),
          (39, 1),
          (50, 1),
          (52, 2),
          (71, 1),
          (99, 1),
          (273, 1),
          (362, 1),
```

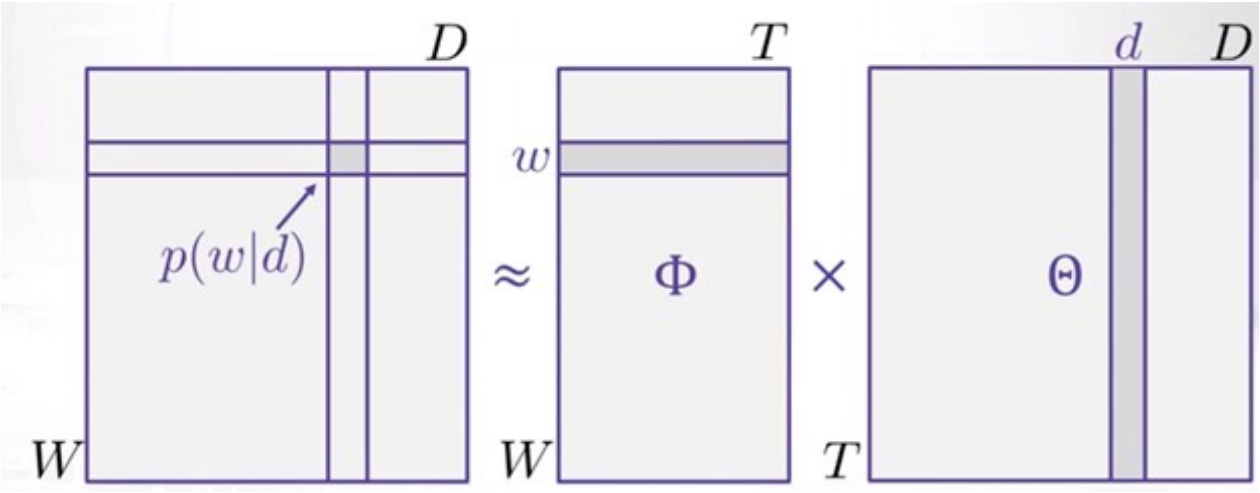
4. Construction du modèle LDA

lda_model = gensim.models.ldamodel.LdaModel

- corpus=corpus
- Id2word=id2word
- num_topics=
- random_state=1,
- passes=10,
- per_word_topics=True



Matrice correspondant au calcul et à l'assignation des *topics*



Les topics sont créés à l'aide des fréquences de mots dans les documents (bag-of-words) où l'on calcule la probabilité qu'un mot soit inclus dans un topic et qu'un topic soit inclus dans un document. L'itération de ces distributions aléatoires produits des topics de mots apparaissant fréquemment ensemble.

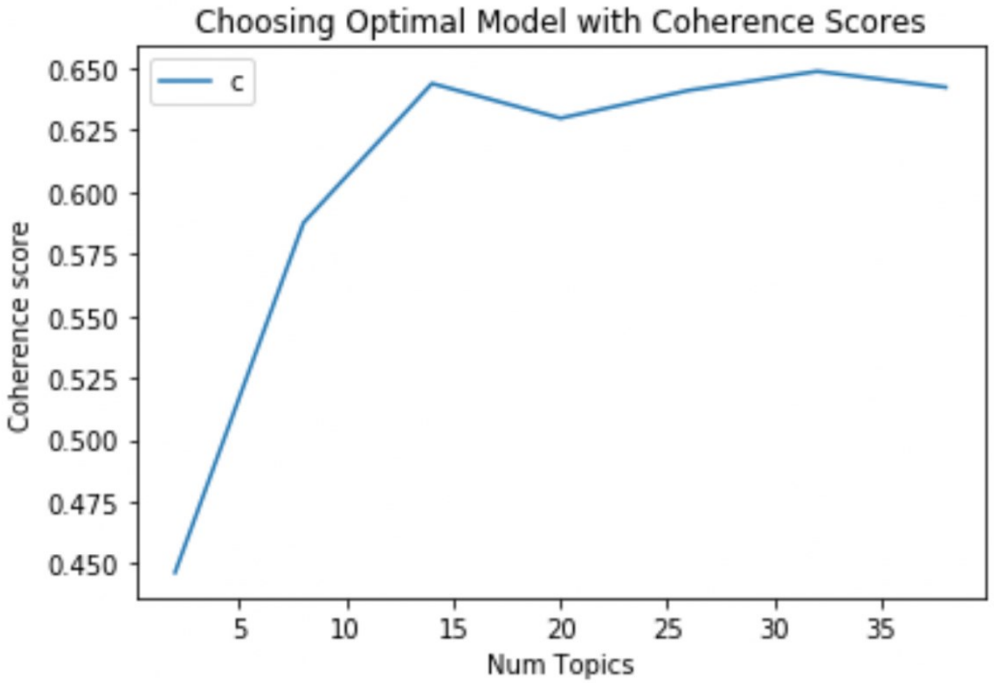
Donc, un même mot peut servir à plusieurs topics différents s'il est utilisé dans différents contextes.

Source : https://medium.com/@tomar.ankur287/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045

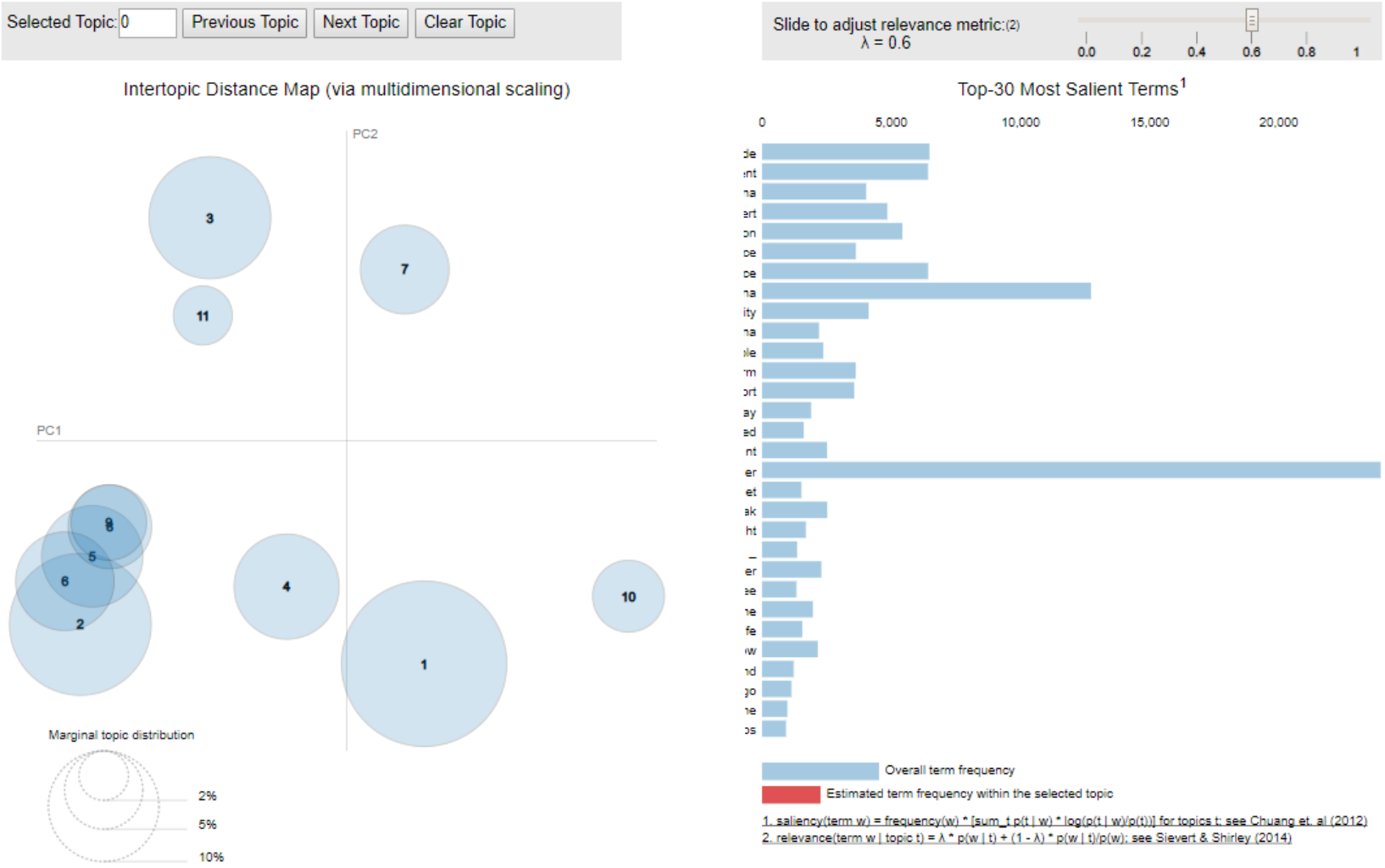
5. Évaluation du modèle

Calculer le nombre de *topics* (k) optimal en vérifiant le score de cohérence (c).

- Généralement, le nombre de topics optimal selon cette mesure correspond au moins grand nombre de topics avant que le score de cohérence ne se mette à baisser.



6. Visualisation



https://gardeso.com/

Maxime Bérubé, PhD
Chercheur postdoctoral
Concordia University/University of Waterloo
maxime.berube@concordia.ca