Challenge Développement Durable

...

Constanza Bello
Laura Abu El Ghait
Léandre Bischoff
Rayan Merkhi

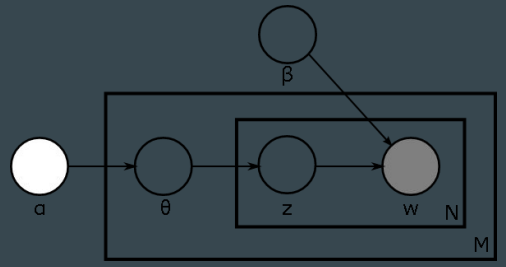
Enjeu : Identifier les cibles des ODD 12, 15 et 16 dans un texte d'une manière pouvant être généralisée aux 169 cibles des 17 ODD

<u>Notre solution</u> : Utiliser un topic modeling non supervisé et semi supervisé avec LDA et BERTopic

BERTopic

BERTopic est une méthode de modélisation de sujets utilisant des transformateurs et un c-TF-IDF créant ainsi des clusters denses. Cette technique a l'avantage de permettre la visualisation facile des sujets.

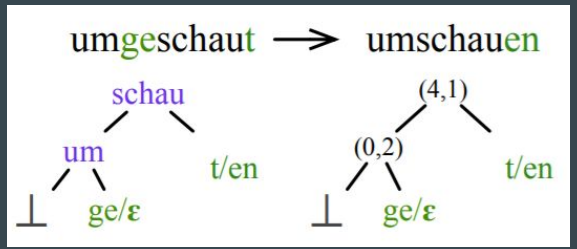
LDA



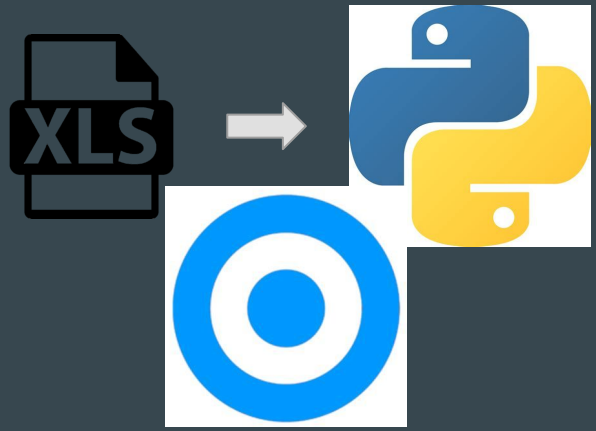
L'algorithme prend une base de données de phrases et crée ses propres groupes en assimilant aux mots clés un des groupes et vérifie le groupe le plus présent dans la phrase afin de le lui assigner.

Processus BERTopic:

1. Extraction du texte d'un pdf en format texte
2. Importation des données dans Python
3. Établissement de la liste des cibles
4. Processing des données et suppression des stop words
5. Lemmatisation
6. Seed words
7. Traitement des données
8. Sortie et visualisation des données



Processus LDA:



1. Extraction du texte d’un xls
2. Importation des données dans Python
3. Déclaration du nombre de cibles voulues
4. Processing des données et suppression des stop words
5. Filtrage à l'aide d'un dictionnaire
6. Traitement des données
7. Sortie et visualisation des données

```
Taux de réussite total : 55%
```

Avantages :

★ Implémentation rapide avec les bon outils

BERTopic:

★ Solution multilingue
★ Seed words
★ Visualisation

LDA:

★ Résultat aisément globalisable sur les 169 cibles
★ Améliorations envisageables:
  - Active Training
  - Boucle de rétroaction
  - Restricted LDA

Piste peu explorée malgré son fort potentiel

Inconvénients:

LDA :

- Difficulté d'implémentation actuelle
- Nécessité d’outils qui ne sont pas au point

BERTopic:

- Besoin d’une base de données conséquente

Merci pour votre attention !

Sources:

https://aclanthology.org/W14-3104.pdf
https://www.cs.uic.edu/~liub/publications/PAKDD-2011.pdf
https://github.com/yya518/sparse-constrained-lda
https://github.com/MaartenGr/BERTopic