



Social Computing & Analyse de Communautés

Matthieu François

March 28, 2022

References I

-  Igor Brigadir, Derek Greene, and Padraig Cunningham, *Analyzing discourse communities with distributional semantic models*, 07 2015.
-  Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg, *Processing social media messages in mass emergency: A survey*, *ACM Computing Surveys (CSUR)* **47** (2015), no. 4, 1–38.

Contexte

- Détection d'évènement

Contexte

- Détection d'évènement
- Détection de communauté

Contexte

- Détection d'évènement
- Détection de communauté
- Clustering

Contexte

- Détection d'évènement
- Détection de communauté
- Clustering

Nécesite

- diversifier les sources de données
- vérifier leur crédibilité

Contexte

Analyzing discourse communities with distributional semantic models

Nous nous plaçons dans une contexte d'analyse du discours critique (CDA), dans lequel l'analyse des fréquences de termes est fréquemment utilisée.

Contexte

Analyzing discourse communities with distributional semantic models

Nous nous plaçons dans une contexte d'analyse du discours critique (CDA), dans lequel l'analyse des fréquences de termes est fréquemment utilisée.

Limites

- Comment se baser sur des fréquences lorsque la sémantique d'un mot peut changer ?

Contexte

Analyzing discourse communities with distributional semantic models

Nous nous plaçons dans une contexte d'analyse du discours critique (CDA), dans lequel l'analyse des fréquences de termes est fréquemment utilisée.

Limites

- Comment se baser sur des fréquences lorsque la sémantique d'un mot peut changer ?
 - ☞ Identification du contexte (ex : KeyWord In Context)

Contexte

Analyzing discourse communities with distributional semantic models

Nous nous plaçons dans une contexte d'analyse du discours critique (CDA), dans lequel l'analyse des fréquences de termes est fréquemment utilisée.

Limites

- Comment se baser sur des fréquences lorsque la sémantique d'un mot peut changer ?
 - ☞ Identification du contexte (ex : KeyWord In Context)
- Données trop denses pour faire du KWIC

Contexte

Analyzing discourse communities with distributional semantic models

Nous nous plaçons dans une contexte d'analyse du discours critique (CDA), dans lequel l'analyse des fréquences de termes est fréquemment utilisée.

Limites

- Comment se baser sur des fréquences lorsque la sémantique d'un mot peut changer ?
 - ☞ Identification du contexte (ex : KeyWord In Context)
- Données trop denses pour faire du KWIC
 - ☞ Utilisation de méthodes non supervisées

Analyse non supervisée de documents

Lignes de concordance et KWIC

Ex : Les lignes de concordance et les KWIC

Méthodes d'indexation de documents / textes : centre chaque occurrence du terme recherché, permettant de voir les mots avant et après afin d'établir un contexte

Analyse non supervisée de documents

Lignes de concordance et KWIC

Ex : Les lignes de concordance et les KWIC

Méthodes d'indexation de documents / textes : centre chaque occurrence du terme recherché, permettant de voir les mots avant et après afin d'établir un contexte

Méthodes classiques (ou non) reposent sur l'hypothèse de distributivité.

Analyse non supervisée de documents

Lignes de concordance et KWIC

Ex : Les lignes de concordance et les KWIC

Méthodes d'indexation de documents / textes : centre chaque occurrence du terme recherché, permettant de voir les mots avant et après afin d'établir un contexte

Méthodes classiques (ou non) reposent sur l'hypothèse de distributivité.

Hypothèse de distributivité

Les mots de mêmes sens sont localement proches dans une phrase.

Context	Keyword ^	Context
virgin engaged to a man whose name was	Joseph	, of the house of David. The virgin's
to their own towns to be registered. 4	Joseph	also went from the town of Nazareth in
they went with haste and found Mary and	Joseph	, and the child lying in the manger. 17
of Jacob, 16 and Jacob the father of	Joseph	the husband of Mary, of whom Jesus was
When his mother Mary had been engaged to	Joseph	, but before they lived together, she was found
child from the Holy Spirit. 19 Her husband	Joseph	, being a righteous man and unwilling to expose
appeared to him in a dream and said, "	Joseph	, son of David, do not be afraid to
which means, "God is with us." 24 When	Joseph	awoke from sleep, he did as the angel
left, an angel of the Lord appeared to	Joseph	in a dream and said, "Get up, take
for the child, to destroy him." 14 Then	Joseph	got up, took the child and his mother

Analyse non supervisée de documents

Topic Modeling

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)

Analyse non supervisée de documents

Topic Modeling

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)
 - Ne prend pas en compte la sémantique des mots

Analyse non supervisée de documents

Topic Modeling

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)
 - Ne prend pas en compte la sémantique des mots

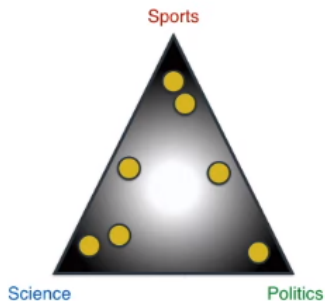
Attribution de X topics à des documents par processus itératif selon la loi de Dirichlet.

Latent Dirichlet Allocation

$$P(W, Z, \theta, \phi, \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\phi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi_{Z_{j,t}}) \quad (1)$$

Latent Dirichlet Allocation

$$P(W, Z, \theta, \phi, \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\phi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi_{Z_{j,t}}) \quad (1)$$



Loi de Dirichlet

Densité :

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_1^K x_i^{\alpha_i - 1} \quad (2)$$

Word embedding et Analyse de communautés

Contexte

Dans *Analyzing discourse communities with distributional semantic models*, Brigadir cherche à :

- Comprendre l'évolution des débats / pensées lors du référendum sur l'indépendance de l'Ecosse (2014)
- Comprendre l'évolution des débats / pensées lors des élections de mi-mandat aux US (2014)

Word embedding et Analyse de communautés

Contexte

Dans *Analyzing discourse communities with distributional semantic models*, Brigadir cherche à :

- Comprendre l'évolution des débats / pensées lors du référendum sur l'indépendance de l'Ecosse (2014)
- Comprendre l'évolution des débats / pensées lors des élections de mi-mandat aux US (2014)

<i>Community</i>	<i>Users</i>	<i>Tweets</i>	<i>Total Words</i>	<i>Date Range</i>
Scotland Yes	618	799,096	12,551,654	11-Aug to 19-Oct
Scotland No	610	570,024	8,957,721	11-Aug to 19-Oct
Democrat	942	89,296	1,404,737	10-Oct to 20-Nov
Republican	997	80,840	1,209,197	10-Oct to 20-Nov

Word embedding et Analyse de communautés

Contexte

Dans *Analyzing discourse communities with distributional semantic models*, Brigadir cherche à :

- Comprendre l'évolution des débats / pensées lors du référendum sur l'indépendance de l'Ecosse (2014)
- Comprendre l'évolution des débats / pensées lors des élections de mi-mandat aux US (2014)

<i>Community</i>	<i>Users</i>	<i>Tweets</i>	<i>Total Words</i>	<i>Date Range</i>
Scotland Yes	618	799,096	12,551,654	11-Aug to 19-Oct
Scotland No	610	570,024	8,957,721	11-Aug to 19-Oct
Democrat	942	89,296	1,404,737	10-Oct to 20-Nov
Republican	997	80,840	1,209,197	10-Oct to 20-Nov

Approche

- Fenetrage par semaine afin de mesurer l'évolution
- Elaboration des communautés selon la une liste d'officiels, la description de users / leur image.

Méthode

Word2Vec

- CBOW
Prediction d'un terme à partir des mots de contexte environnant

Méthode

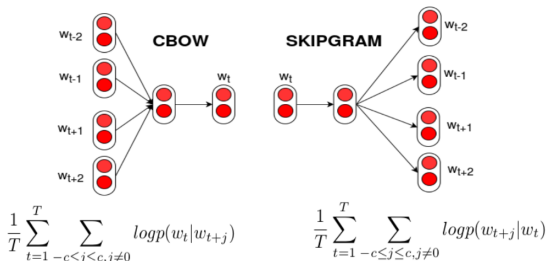
Word2Vec

- CBOW
Prediction d'un terme à partir des mots de contexte environnant
- skipgram
Prediction des mots de contexte environnant à partir d'un terme

Méthode

Word2Vec

- CBOW
Prediction d'un terme à partir des mots de contexte environnant
- skipgram
Prediction des mots de contexte environnant à partir d'un terme



Méthode

Word2Vec

- skipgram
Identifiable au KWIC

Méthode

Score Jaccard

Objectif : Mesurer la similarité de terme au sein d'une communauté selon le temps ou entre 2 Communautés

Méthode

Score Jaccard

Objectif : Mesurer la similarité de terme au sein d'une communauté selon le temps ou entre 2 Communautés

- On souhaite comparer plusieurs espaces entre eux

Méthode

Score Jaccard

Objectif : Mesurer la similarité de terme au sein d'une communauté selon le temps ou entre 2 Communautés

- On souhaite comparer plusieurs espaces entre eux
- Les comparer numériquement n'est pas possible car construit indépendamment

Méthode

Score Jaccard

Formule de Jaccard

On se base sur les cardinaux de chaque terme

$$AJ(A, B) = \frac{1}{k} \sum_{d=1}^k J_d(A, B) \quad (3)$$

Méthode

Score Jaccard

Formule de Jaccard

On se base sur les cardinaux de chaque terme

$$AJ(A, B) = \frac{1}{k} \sum_{d=1}^k J_d(A, B) \quad (3)$$

Plus le contexte sera similaire pour 1 mot clé de référence, plus le score moyen sera proche de 1. Cette mesure ne nécessite pas de comprendre le contexte. Cela se fait dans un second temps.

Méthode

Résumé

Fenetrage par semaine

- Entraînement d'un embedding / semaine / communauté
- Moyenne des scores Jaccard de 1 à k voisins

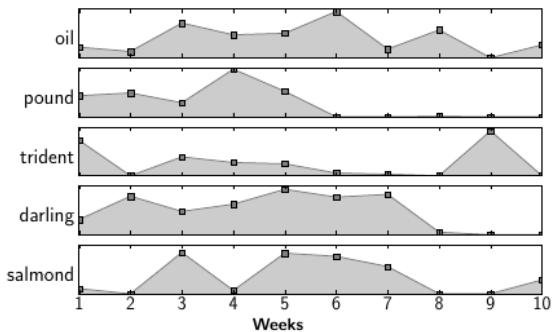
Méthode

Paramètres

- $K = 30$
- skipgram window = 5
- embedding $n = 300$

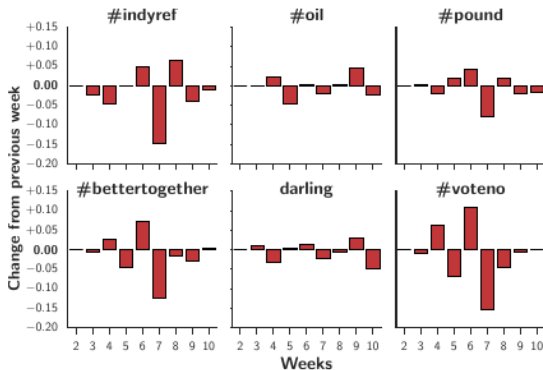
Résultats

Evolution du score Jaccard moyen selon le temps



Résultats

Evolution du score Jaccard moyen selon les non



(b) Selected “No” community word neighborhoods.

Perspectives

Etudier l'importance donnée à l'écologie dans le discours des politiques sur twitter avant les élections présidentielles.

Perspectives

Etudier l'importance donnée à l'écologie dans le discours des politiques sur twitter avant les élections présidentielles.

- 8.565 comptes twitter reliés à des politiques

Perspectives

Etudier l'importance donnée à l'écologie dans le discours des politiques sur twitter avant les élections présidentielles.

- 8.565 comptes twitter reliés à des politiques
- +5 millions de tweets collectés

Perspectives

Mots clés potentiels

- nucléaire
- énergie
- déchets
- climat
- carbone

Christel W. van Eck, 2020, Climate Change Risk Perceptions of Audiences in the Climate Change Blogosphere

Modélisation de la perception des risques du changement climatique par différentes variables : les émotions / les normes sociales / la confiance / les connaissances / les expériences
A partir de réponses sur une échelle de 1 à 8 de questions telles que :

- "Selon vous, quelle est la gravité de la menace que représente le changement climatique pour l'environnement naturel ?"
- "Je crois que le changement climatique est quelque chose de très positif"

[ICDV15] [BGC15]