Métriques et Statistiques en TALN

A. Pappa

March 18, 2024

Table de matières

Introduction aux Métriques en TALN

- Importance des métriques pour évaluer l'efficacité des modèles.
- Vue d'ensemble des types de métriques: basées sur l'erreur, basées sur la similarité.

Table de matières

TF-IDF (Term Frequency-Inverse Document Frequency)

- **Définition**: Mesure de l'importance d'un terme dans un document par rapport à un corpus.
- **Formule**: $TF - IDF = TF(t, d) * IDF(t, D)$
  - TF(t, d) = (Nombre de fois que le terme t apparaît dans un document d) / (Nombre total de termes dans le document d)
  - IDF(t, D) = log_e(Total de documents dans le corpus D / Nombre de documents contenant le terme t)
- **Application** : Recherche d'information, extraction de caractéristiques pour la classification de textes.
- **Exemple** : Calcul de TF-IDF pour un terme dans différents documents.

TF-IDF exemple

Contexte :
Supposons que nous avons un corpus de documents sur différents sujets. Nous voulons déterminer l'importance du terme *intelligence* dans un document spécifique sur l'intelligence artificielle.

Exemple sur TF-IDF :

- Document d: *L'intelligence artificielle transforme notre manière de comprendre l'intelligence humaine.*
- Corpus D: 100 documents, dont 5 contiennent le mot *intelligence*.

Exemple suite ...

Calcul:

- TF(*intelligence*, d) = 2/10 (le terme *intelligence* apparaît 2 fois sur 10 termes dans le document)
- IDF(*intelligence*, D) = $log(100/5) = log(20)$

$$\textbf{TF-IDF}(intelligence,\ d, D) = TF * IDF = (2/10) * log(20)$$

Le TF-IDF élevé indique que le terme *intelligence* est important dans le contexte du document donné au sein du corpus.

Table de matières

Mesures de Similarité

- **Objectif**: Évaluer à quel point deux textes ou documents sont similaires.
- **Exemples de mesures**:
  - Distance de Jaccard
  - Similarité cosinus
- **Application** :Systèmes de recommandation, détection de plagiat.

Exemple de similarité cosinus

Similarité cosinus entre deux documents

Documents :

- Document A: *Le chat aime les croquettes.*
- Document B: *Le chien aime les croquettes.*

Exemple similarité cosinus suite ...

Vecteurs (basés sur un simple modèle sac de mots) :

- $A = [1, 0, 1, 1]$ (pour "le", "chat", "aime", "croquettes")
- $B = [1, 1, 1, 1]$ (pour "le", "chien", "aime", "croquettes")

Calcul de la similarité cosinus : La similarité cosinus mesure l'angle entre les deux vecteurs. Un angle plus petit indique une plus grande similarité.

$$\textbf{Similarité cosinus} = (A.B)/(||A|| * ||B||)$$

La similarité cosinus serait relativement élevée, reflétant que les deux documents parlent de sujets similaires (animaux domestiques et leurs préférences alimentaires).

Exemple similarité Jaccard

Définition :
La **similarité de Jaccard** mesure à quel point deux ensembles sont similaires.
Elle est définie comme la taille de l'intersection divisée par la taille de l'union des deux ensembles.
Formellement, pour deux ensembles A et B la similarité de Jaccard

$$J(A, B)J(A, B)$$

est calculée comme suit :

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Exemple similarité Jaccard suite ...

Considérons deux phrases pour simplifier :

- Phrase **A** : *Le chat dort sur le tapis.*
- Phrase **B**: *Le chien dort sur le tapis.*

Exemple similarité Jaccard suite ...

Pour calculer **similarité de Jaccard** entre ces deux phrases, transformons d'abord chaque phrase en un ensemble de mots (en ignorant la ponctuation et en considérant chaque mot une seule fois) :

- Ensemble **A** = "Le", "chat", "dort", "sur", "le", "tapis" = "Le", "chat", "dort", "sur", "tapis"
- Ensemble **B** = "Le", "chien", "dort", "sur", "le", "tapis" = "Le", "chien", "dort", "sur", "tapis"

Exemple similarité Jaccard suite ...

Maintenant, calculons l'intersection et l'union de ces deux ensembles :

$$\textbf{Intersection}(A \cap B) = \{"\text{Le}", "\text{dort}", "\text{sur}", "\text{tapis}"\}$$

$$\textbf{Union}(A \cup B) = \{"\text{Le}", "\text{chat}", "\text{dort}", "\text{sur}", "\text{tapis}", "\text{chien}"\}$$

Exemple similarité Jaccard suite ...

La taille de l'intersection est **4**, et la taille de l'union est **6**.
Ainsi, la **similarité de Jaccard** entre la **Phrase A** et la **Phrase B** est :

$$J(A, B) = \frac{4}{6} = \frac{2}{3} \approx 0.67$$

Ce résultat indique que les deux phrases sont assez similaires selon la mesure de Jaccard, avec une **similarité** d'environ **67%**.
Cet exemple illustre comment la **similarité de Jaccard** peut être utilisée pour quantifier la similitude entre deux ensembles de données, comme des phrases dans ce cas.

Précision, Rappel et F1-Score

- **Définitions** :
  - **Précision**: Proportion des identifications positives qui sont effectivement correctes.
  - **Rappel**: Proportion des vrais positifs qui ont été correctement identifiés.
  - **F1-Score** : Moyenne harmonique de la précision et du rappel.
- **Formules** :
  1. **Précision** $= TP/(TP + FP)$
  2. **Rappel** $= TP/(TP + FN)$
  3. **F1-Score** $= 2*$ (Précision $*$ Rappel) / (Précision + Rappel)
- **Application** : Évaluation des systèmes de classification, en particulier lorsque les classes sont déséquilibrées.
- **Exemple** : Calcul de ces métriques pour un système de classification de textes.

Exemple de classification

- **Contexte** : Un système de classification des emails en *spam* et *non-spam*.
- **Données** :
  - 100 emails à classifier, dont 20 sont des spams.
  - Le système identifie correctement 15 spams (vrais positifs) et marque 5 non-spams comme spams (faux positifs).
  - 5 spams sont manqués (faux négatifs).
- **Calculs** :
  1. **Précision** $= \frac{TP}{TP+FP}$[1][2] $= \frac{15}{15+5} = 0.75$
  2. **Rappel** $= \frac{TP}{TP+FN} = \frac{15}{15+5} = 0.75$
  3. **F1-Score** $= 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} = 2 \times \frac{0.75 \times 0.75}{0.75+0.75} = 0.75$

[1] TP (True Positives)
[2] FP (False Positives)

Table de matières

Conclusion

- Importance de choisir la bonne métrique en fonction du problème et des données.
- Réflexion sur les limitations de chaque métrique et comment les combiner pour une évaluation complète.