Recherche d'Information

Pondération

Dr. Yassine Drias
Université d'Alger I Benyoucef Benkhedda

Termes importants

- Définir l'unité d'indexation (radical, lemme, mot simple, groupe de mots) permet de trouver les termes importants caractérisant le contenu de chaque document

- Le **pouvoir de caractérisation** de ces termes dépend de plusieurs facteurs
  - Certains termes sont plus importants que d'autres dans la caractérisation du contenu

Déterminer les termes importants dans un document

- La pondération des termes
  - Mesure l'importance d'un terme dans un document et/ou dans un corpus
  - Comment représenter au mieux le contenu d'un document ?
- Considérations statistiques, parfois linguistiques
  - Loi de Zipf: élimination des termes trop fréquents ou trop rares
- Facteurs de pondération
  - e.g. pondération locale par rapport au document, pondération globale par rapport à une collection de documents
  - Normalisation: prise en compte de la longueur des documents, etc
- Les termes importants doivent avoir un poids fort

Approches de pondération

- La méthode de pondération peut dépendre du modèle de recherche d'information.
- Plusieurs approches de pondération:
  - TF-IDF approche plus répandue
  - Pourvoir discriminatoire d'un terme
  - Modèle 2-poisson
  - Clumping model
  - Modèle de Langage

Une méthode de référence TF-IDF

- Le TF-IDF est une méthode de pondération souvent utilisée en recherche d'information.
- Mesure statistique qui permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus.
  - Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document.
  - Il varie également en fonction de la fréquence du mot dans le corpus.

TF – Pondération locale

- TF (Term Frequency) estime l'importance d'un terme dans un document en se basant sur la fréquence d'apparition du terme dans ce document
  - Plus un terme est fréquent dans un document plus il est important dans la description de ce dernier
  - $fréq_{ij}$: nombre d'occurrences du terme t_i dans le document d_j
- Plusieurs variantes ont été proposées

Calcul de TF - variantes

- Binaire $$tf_{ij} = \begin{cases} 1 & si\ t_i \epsilon\ d_j \\ 0 & sinon \end{cases}$$
- Fréquence brute $$tf_{ij} = fréq_{ij}$$
- Normalisation logarithmique $$tf_{ij} = 1 + \log(fréq_{ij})$$
- Normalisation par le max $$tf_{ij} = \frac{fréq_{ij}}{max_{\{\forall\ t_l \in d_j\}}\ fréq_{lj}}$$

Calcul de TF - variantes

- Normalisation par taille du doc $$tf_{ij} = \frac{fréq_{ij}}{\sum_{\forall\, t_l \in d_j} fréq_{lj}}$$

- Okapi TF (Robertson) $$tf_{ij} = \frac{fréq_{ij}}{fréq_{ij} + k_1.(1 - b + b * \frac{|d_j|}{avgdl})}$$
  - k_1, b: paramètres libres, généralement:
    - $k_1 \in [1{,}2\ ;\ 2{,}0]$ et $b = 0{,}75$
  - $|d_j|$: longueur du document (en nombre de termes)
  - avgdl: longueur moyenne des documents du corpus

IDF – Pondération globale

- Un terme a un pouvoir de caractérisation important si sa fréquence est **élevée dans un document** du corpus et **basse dans le reste des documents**.
- IDF (Inverse Document Frequency) estime l'importance du terme par rapport à sa fréquence d'apparition dans une collection de documents

$$\log\left(\frac{N}{n_i}\right) \quad \text{ou} \quad \log\left(\frac{N}{n_i}+1\right)$$

N : le nombre de documents de la collection
n_i: le nombre de documents contenant le terme t_i

TF * IDF

- $w(t_i, d_j)$: poids du terme t_i dans le document d_j, on utilise aussi la notation w_{ij}

1. $$w(t_i, d_j) = \frac{fréq_{ij}}{max_{\{\forall\, t_l \in d_j\}}\, fréq_{lj}} * \log\left(\frac{N}{n_i} + 1\right)$$

2. $$w(t_i, d_j) = 1 + \log(fréq_{ij}) * \log\left(\frac{N}{n_i} + 1\right)$$

3. $$w(t_i, d_j) = \frac{1+\log(fréq_{ij}) * \log\left(\frac{N}{n_i}\right)}{\sum_{\forall\, t_l \epsilon d_j} 1+\log(fréq_{lj}) * \log\left(\frac{N}{n_l}\right)}$$

Exploitation en RI

- En recherche d'information, la mesure TF-IDF est utilisée pour déterminer le score d'un document vis-à-vis de la requête.
- Soit une requête $q(t_1, t_2)$ et un document $d_1(t_1, t_2, ..t_n)$
  - Une manière de calculer le score du document par rapport à la requête est de faire la somme pondérée des termes de la requête apparaissant dans le document

$$score(q, d_1) = \sum_{\forall\, t_i \epsilon q,\, t_i \epsilon d_1} w(t_i, d_1)$$

Valeur discriminatoire d’un terme

- Modèle de pondération proposé par Salton, 1975

- Valeur discriminatoire d’un terme
  - Mesure la capacité d’un terme dans la distinction de documents.
  - Consiste à comparer la similarité entre documents avec et sans ce terme.

Valeur discriminatoire d'un terme

- Calculer la densité de l'espace de documents

$$avgsim = K * \sum_{i=1}^{n} \sum_{i \neq j, j=1}^{n} sim(d_i, d_j)$$

  - $K = \frac{1}{n(n-1)}$

- La valeur discriminatoire d'un terme est :

$$Disc(t_k) = avgsim\{-t_k\} - avgsim$$

  - Un bon terme à une valeur positive
  - Un terme indifférent à une valeur proche de zéro
  - Un terme pauvre a une valeur négative

La loi de Zipf

- George Kingsley Zipf (1902-1950) est un linguiste et philologue américain qui étudia la statistique appliquée aux différentes langues.
- La loi de Zipf est une observation **empirique** concernant la **fréquence** des mots dans un texte
- Le principe du moindre effort (least effort)
  - Il est plus simple pour un auteur (rédacteur d'un document) de répéter les mots que d'en chercher de nouveaux.
- "Term frequency decreases very rapidly with the rank"

Observation de Zipf

- Si on classe les mots dans l'ordre décroissant de leur fréquence, et on leur attribue un numéro de rang, alors:
  - Rang * fréquence ≈ constante
  - $r * p_r = A$
    - r : le rang du terme basé sur sa fréquence
    - $p_r = \frac{fréquence\ du\ terme\ de\ rang\ r}{N}$
    - N: nombre total d'occurrence de tous les termes
    - A ≈ 0.1

Rank	Word	Number of occurrences
1	the	69975
2	of	36432
3	and	28872
4	a	26800
5	to	26190
6	in	21338
7	he	20033
8	have	12458
9	it	11247
10	that	10790

Exemple de la loi de zipf

Word	Freq	r	Pr(%)	r*Pr
the	2,420,778	1	6.488	0.0649
of	1,045,733	2	2.803	0.0561
to	968,882	3	2.597	0.0779
a	892,429	4	2.392	0.0957
and	865,644	5	2.32	0.116
in	847,825	6	2.272	0.1363
said	504,593	7	1.352	0.0947
for	363,865	8	0.975	0.078
that	347,072	9	0.93	0.0837
was	293,027	10	0.785	0.0785
on	291,947	11	0.783	0.0861
he	250,919	12	0.673	0.0807
is	245,843	13	0.659	0.0857
with	223,846	14	0.6	0.084
at	210,064	15	0.563	0.0845
by	209,586	16	0.562	0.0899
it	195,621	17	0.524	0.0891
from	189,451	18	0.508	0.0914
as	181,714	19	0.487	0.0925
be	157,300	20	0.422	0.0843
were	153,913	21	0.413	0.0866
an	152,576	22	0.409	0.09
have	149,749	23	0.401	0.0923
his	142,285	24	0.381	0.0915
but	140,880	25	0.378	0.0944

Word	Freq	r	Pr(%)	r*Pr
has	136,007	26	0.365	0.0948
are	130,322	27	0.349	0.0943
not	127,493	28	0.342	0.0957
who	116,364	29	0.312	0.0904
they	111,024	30	0.298	0.0893
its	111,021	31	0.298	0.0922
had	103,943	32	0.279	0.0892
will	102,949	33	0.276	0.0911
would	99,503	34	0.267	0.0907
about	92,983	35	0.249	0.0872
i	92,005	36	0.247	0.0888
been	88,786	37	0.238	0.0881
this	87,286	38	0.234	0.0889
their	84,638	39	0.227	0.0885
new	83,449	40	0.224	0.0895
or	81,796	41	0.219	0.0899
which	80,385	42	0.215	0.0905
we	80,245	43	0.215	0.0925
more	76,388	44	0.205	0.0901
after	75,165	45	0.201	0.0907
us	72,045	46	0.193	0.0888
percent	71,956	47	0.193	0.0906
up	71,082	48	0.191	0.0915
one	70,266	49	0.188	0.0923
people	68,988	50	0.185	0.0925

Informativité d'un terme