
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

# Analysis and reproducibility of paper "A Word Embedding based Generalized Language Model for Information Retrieval"

D. Ganguly, D. Roy, M. Mitra, G.J.F. Jones

11 february 2020

Alessio Lazzaron, Matteo Marchiori, Andrea Oriolo, Fabio Piras

# Contents

# Introduzione

- Language Model (LM)
  - Sviluppo di modelli probabilistici in grado di prevedere un documento, in una collezione, data una query.

$$P(d|q) = \prod_{t \in q} \lambda \frac{tf(t,d)}{|d|} + (1-\lambda)\frac{cf(t)}{cs}$$

Approcci di LM classici $\begin{cases} \text{Latent Dirichlet allocation (LDA)} \\ \text{Latent semantic analysis (LSA)} \end{cases}$

Si propone un nuovo modello denominato
GENERALIZED LANGUAGE MODEL (GLM)

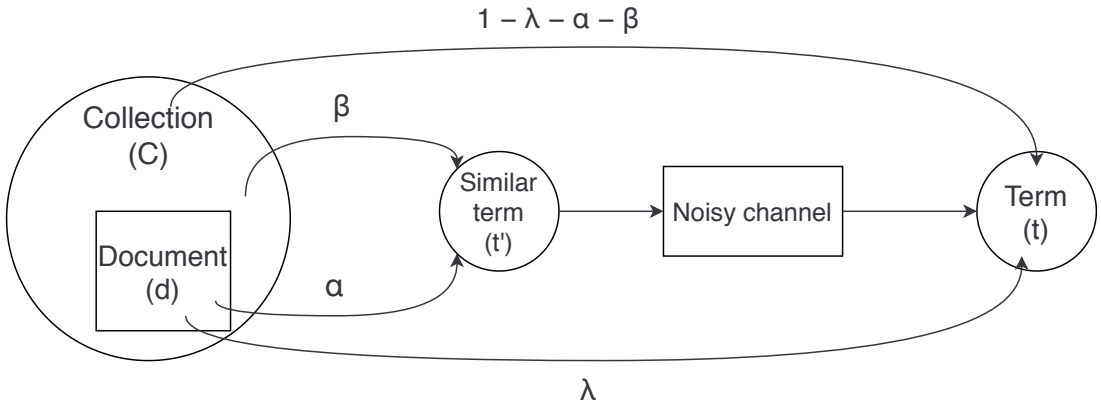
- Word Embedding
  - Rappresentazione delle parole in un nuovo spazio, con lo scopo di memorizzarne le informazioni semantiche e sintattiche.

IDEA

Processo di trasformazione attraverso un «noisy channel» in cui un termine $t'$ viene mutato in un altro termine $t$.

Tre approcci di trasformazione del termine:

- Direct term sampling
- Transformation via Document
- Transformation via Collection



# Formulazione GLM

$$P(t|d) = \lambda P(t|d) + \alpha \sum_{t' \in d} P(t, t'|d) P(t')$$
$$+ \beta \sum_{t' \in N} P(t, t'|C) P(t') + (1 - \lambda - \alpha - \beta) P(t|C)$$

- Termine $\lambda$: Probabilità del termine $t$ della query all'interno del documento $d$ senza trasformazione.
- Termine $\alpha$: Probabilità di trasformare il termine $t$ della query in un termine $t'$ appartenente a $d$.
- Termine $\beta$: Probabilità di trasformare il termine $t$ della query in un termine $t'$ appartenente alla collezione $C$.
- Termine $1 - \lambda - \alpha - \beta$: Probabilità del termine $t$ della query di trovarsi all'interno della collezione $C$.

# Trasformazione di $t$ in $t'$

$$P(t, t'|d) = \frac{sim(t, t')}{\sum(d)} \frac{tf(t', d)}{|d|} \qquad P(t, t'|C) = \frac{sim(t, t')}{\sum(Nt)} \frac{cf(t')}{cs}$$

- Per il calcolo della similarità tra $t$ e $t'$ entrambi i termini vengono rappresentati mediante Word Embedding.
- Una volta ottenuta la loro rappresentazione, per il calcolo dello score di similarità viene utilizzata la Cosine Similarity.

$$sim(A, B) = cos(\vartheta) = \frac{A * B}{\|A\| * \|B\|} = \frac{\sum_{i=1}^{n} Ai * Bi}{\sqrt{\sum_{i=1}^{n} Ai^2} \sqrt{\sum_{i=1}^{n} Bi^2}}$$

# Caratteristiche GLM

$$P(t|d) = \lambda P(t|d) + \alpha \sum_{t' \in d} P(t, t'|d) P(t')$$
$$+ \beta \sum_{t' \in N} P(t, t'|C) P(t') + (1 - \lambda - \alpha - \beta) P(t|C)$$

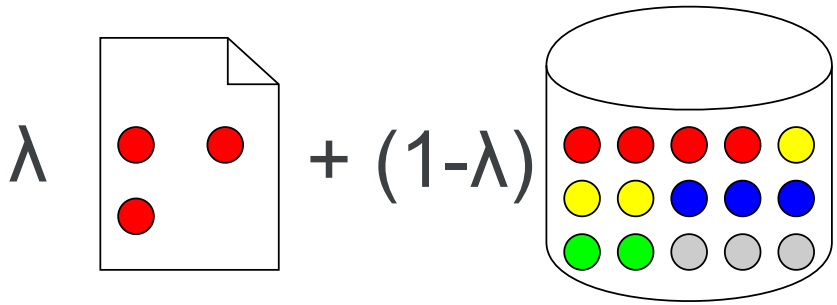
- Maggiore è il valore di similarità tra un termine $t$ della query e i termini $t'$ del documento, maggiore sarà il valore del termine in $\alpha$.
- Maggiore è il valore di similarità tra un termine $t$ della query e i termini $t'$ della collezione, maggiore sarà il valore del termine in $\beta$.
- Il GLM generalizza il Language Model considerando non solo i singoli termini della query ma anche i termini semanticamente simili a essi, favorendo i documenti il cui contesto è simile a quello della query.


GO AHEAD
IM LISTENING
imgflip.com

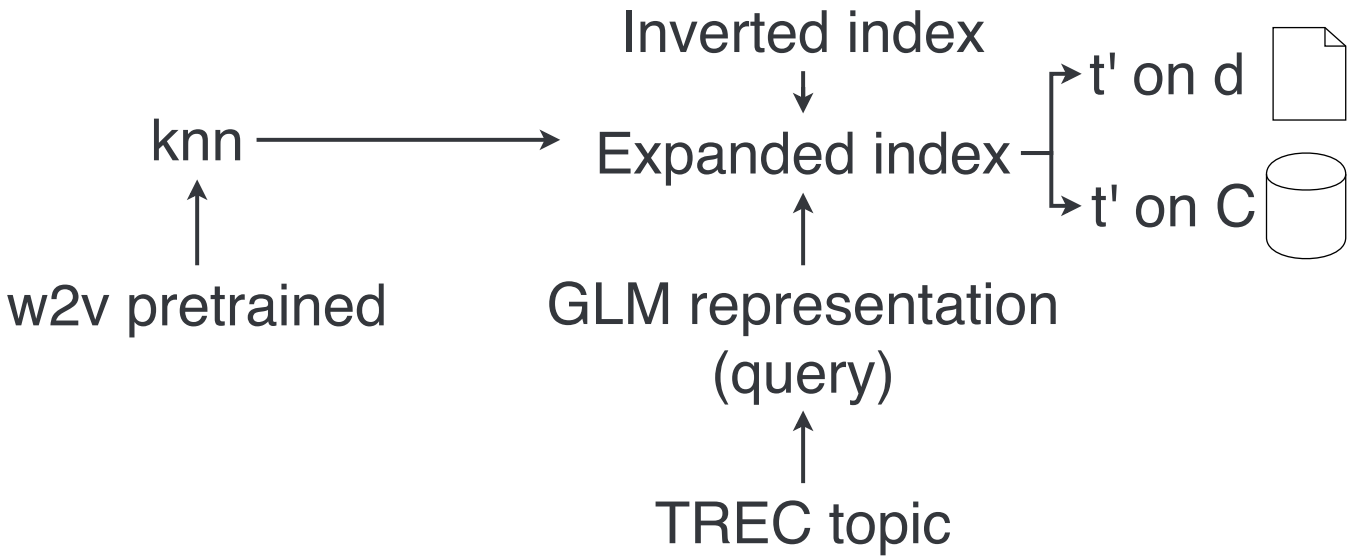
# Indicizzazione e LM

- Lucene... Quale versione?

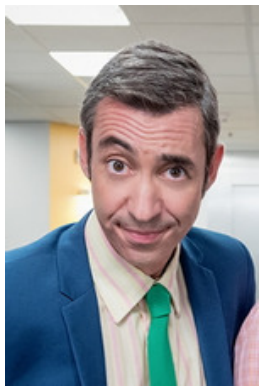

- Indicizzazione... In che modo?
  - Tokenizing
  - Rimozione stopword
  - Indice trasposto
- Base Language Model



# GLM




- Primo approccio: chi fa da sè... Fa da sè!



- Secondo approccio: cerchiamo di prendere esempio e di migliorare
- Vincoli di tempo...

# Risultati

| # | Score | Doc. Id | date | doc_number |
|---|---|---|---|---|
| 0 | 0,6174 | 301084 | 940112 | FT941-16044 |
| 1 | 0,6112 | 129708 | 930226 | FT931-7040 |
| 2 | 0,6112 | 142295 | 930630 | FT932-8 |
| 3 | 0,6112 | 171815 | 930715 | FT933-14247 |
| 4 | 0,6112 | 187823 | 940329 | FT941-706 |
| 5 | 0,6112 | 188903 | 940222 | FT941-7958 |
| 6 | 0,6112 | 297797 | 940127 | FT941-13091 |
| 7 | 0,5239 | 125326 | 930331 | FT931-190 |

La più semplice query con Luke.
'from' è la parola con la frequenza più alta nella collezione.

# Risultati

| Topic Set | MAP | GMAP | RECALL |
|---|---|---|---|
| *TREC-6* | 0.2148<br>0.1612 | 0.0761<br>0.0843 | 0.4778<br>0.5382 |
| *TREC-7* | 0.1771<br>0.1734 | 0.0706<br>0.0843 | 0.4867<br>0.5444 |
| *TREC-8* | 0.2357<br>0.1899 | 0.1316<br>0.0906 | 0.5895<br>0.5484 |
| *Robust* | 0.2555<br>0.2021 | 0.1290<br>0.1016 | 0.7715<br>0.6052 |

Risultati di LM di base

- Indicizzazione con la stoplist standard di Lucene
- Parser standard di Lucene per ottenere query su cui trovare metriche
- Uso di uno stemmer? Improbabile...

# Considerazioni finali

- GLM migliore di LM
- Per ottenere i knn serve molto tempo
- Risultati non riprodotti
- Tempo sottostimato per implementare gli step necessari per riprodurre i risultati del paper
- È stato usato solo Lucene perché è quanto viene usato secondo il paper.

# References


- M. Agosti and G. Silvello, *Slides*, Padua, Italy: academic year 2019-2020.
- D. Ganguly, M. Mitra, D. Roy and G. Jones, *A Word Embedding based Generalized Language Model for Information Retrieval*, Dublin, Ireland; Kolkata, India: SIGIR 2015.
- D. Ganguly, *GLM GitHub repository*
- V. Lavrenko, *Jelinek-Mercer Smoothing video*