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

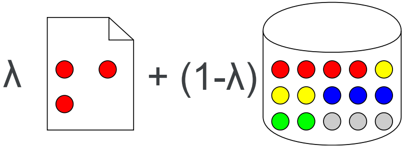
Contents

1 Aspects Reproduced

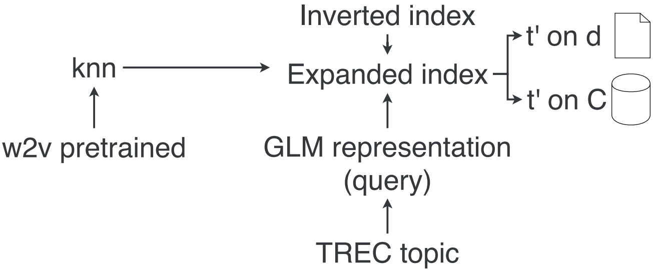
Introduction

Indexing and LM

- Lucene... What version?
- Indexing... How?
  - Tokenization
  - Stopwords removal
  - Transposed index
- Base Language Model



GLM



- First approach: let's do it ourselves (fail)
- Second approach: let's try to take example and make it better
- Is it enough? Time constraints...

Results

Final observations

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*