mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering

Fernando Rezende Zagatti <fernando.zagatti@estudante.ufscar.br>
Paulo Augusto de Lima Medeiros <paulo.medeiros@b2wdigital.com>
Esther da Cunha Soares <esther.soares@b2wdigital.com>
Lucas Nildaimon dos Santos Silva <lucas.nildaimon@estudante.ufscar.br>
Carlos Ramisch <carlos.ramisch@lis-lab.fr>
Livy Real <livy.coelho@b2wdigital.com>

Federal University of São Carlos - Department of Computing, São Carlos, SP, Brazil
Aix Marseille Univ, CNRS, LIS, Marseille, France
americanas s.a. - Front PLN, São Paulo, SP, Brazil

Introduction

Introduction

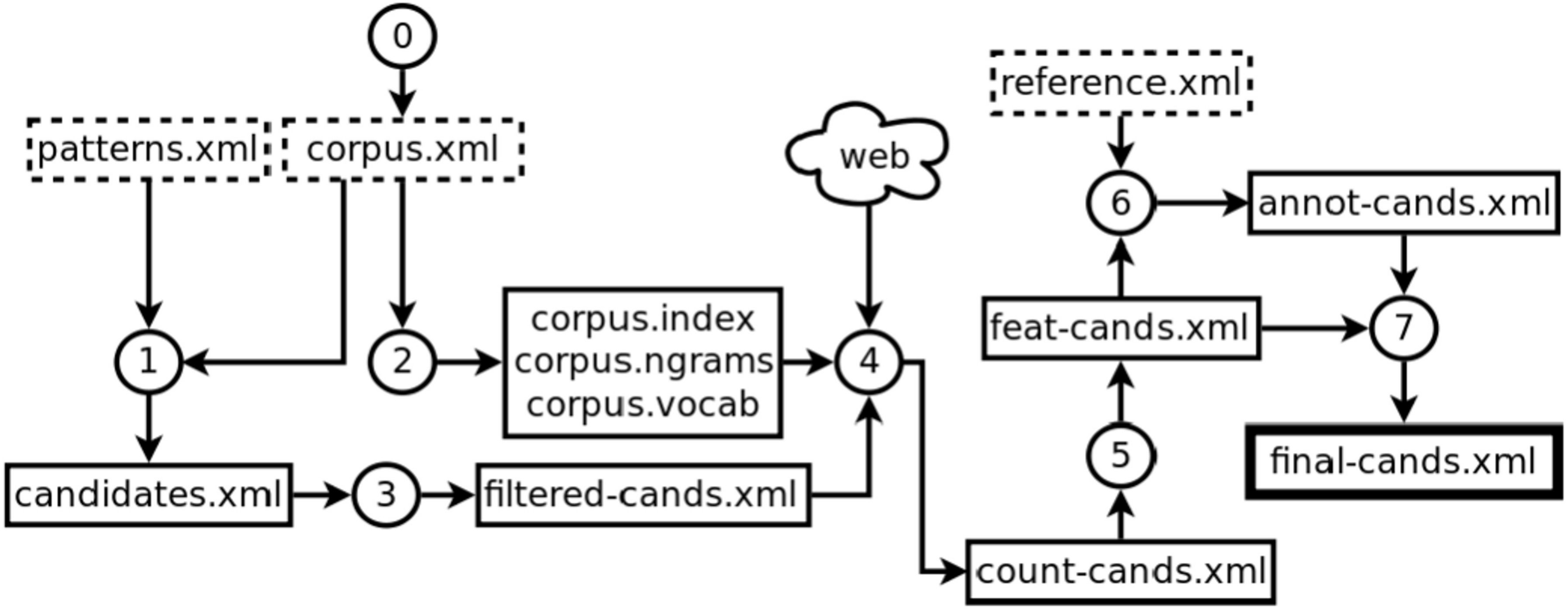
What are Multiword Expressions (MWEs)?

MWEs are combinations of two or more words that present some characteristic behavior when occurring together, having a different behavior when compared to the words used individually.

Examples: Hot dog, Human Resources, Kick the bucket.

Introduction

mwetoolkit



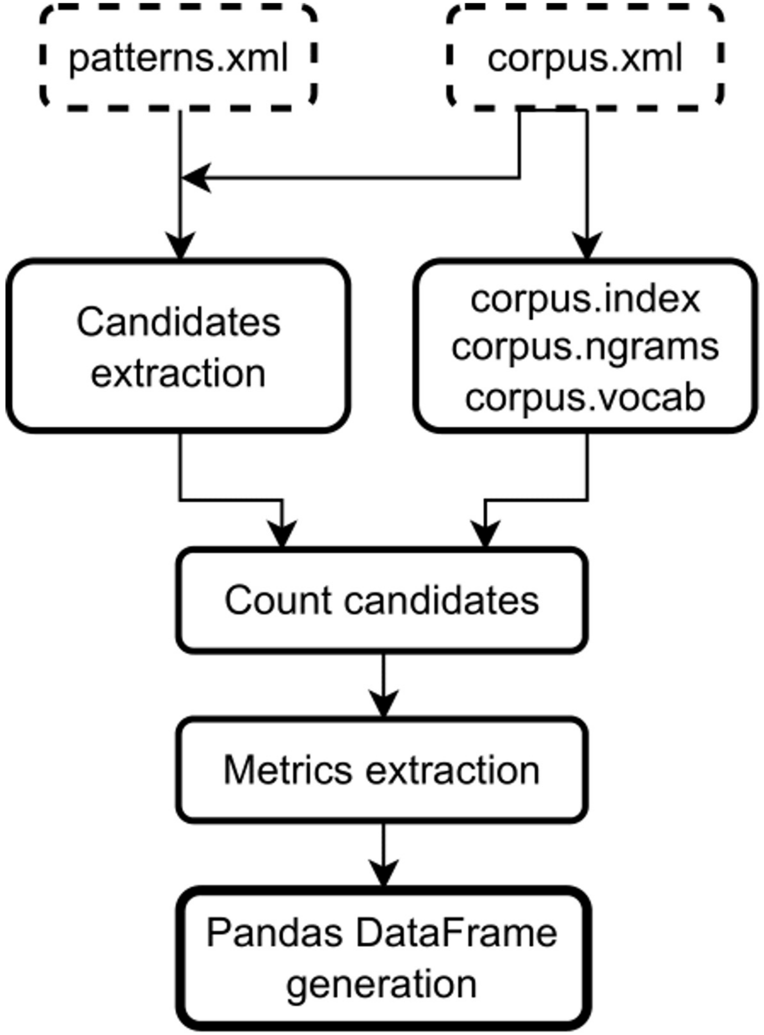
Introduction

What is the problem?

The mwetoolkit is designed to be used through command lines, which makes it difficult to integrate into data science pipelines.

mwetoolkit-lib

mwetoolkit-lib



patterns.xml

corpus.xml

Candidates extraction

corpus.index
corpus.ngrams
corpus.vocab

Count candidates

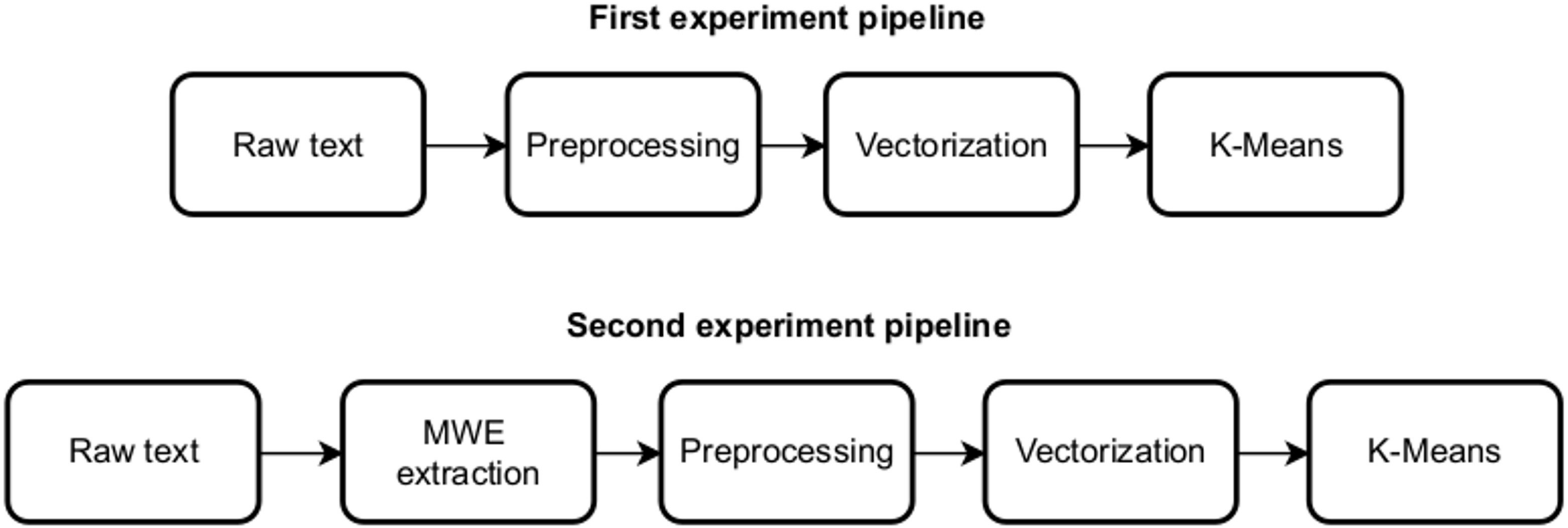
Metrics extraction

Pandas DataFrame generation

Pilot experiment

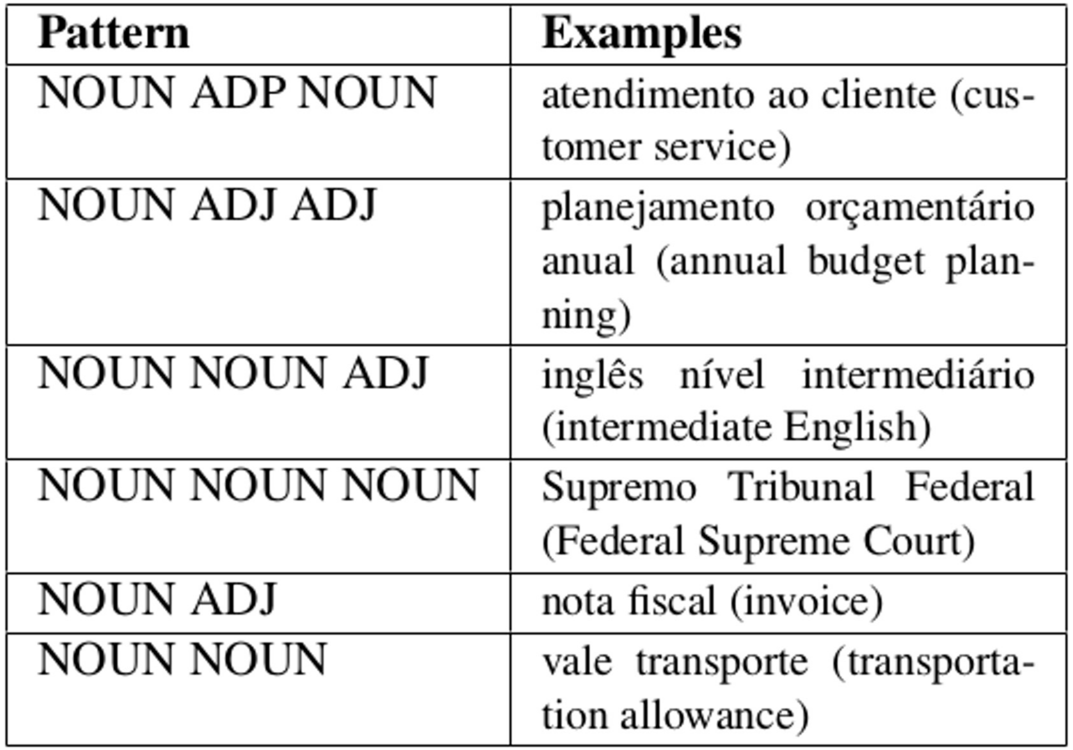
Pilot experiment

- Application to MWE-based Document Clustering
  - Dataset in Portuguese of Human Resources domain from americanas s.a.



Pilot experiment

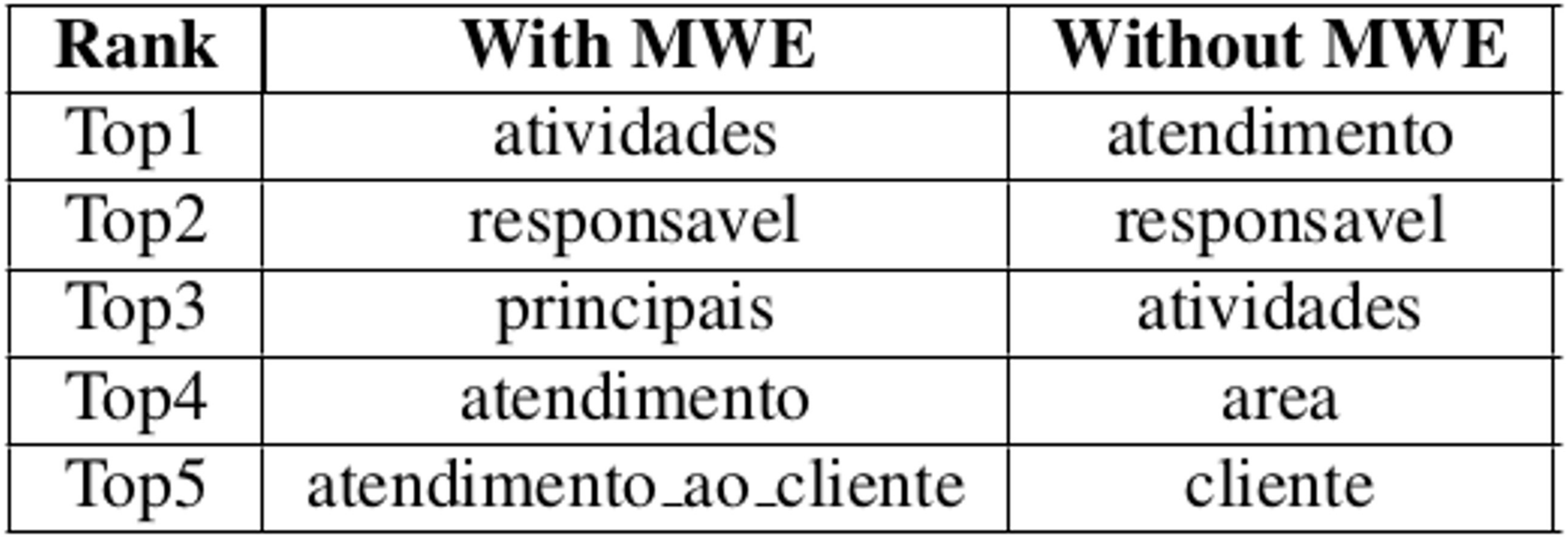
Morphosyntactic patterns used for discovery



Pattern	**Examples**
NOUN ADP NOUN	atendimento ao cliente (customer service)
NOUN ADJ ADJ	planejamento orçamentário anual (annual budget planning)
NOUN NOUN ADJ	inglês nível intermediário (intermediate English)
NOUN NOUN NOUN	Supremo Tribunal Federal (Federal Supreme Court)
NOUN ADJ	nota fiscal (invoice)
NOUN NOUN	vale transporte (transportation allowance)

Pilot experiment

Relevance of words and MWEs by TF-IDF



Rank	With MWE	Without MWE
Top1	atividades	atendimento
Top2	responsavel	responsavel
Top3	principais	atividades
Top4	atendimento	area
Top5	atendimento_ao_cliente	cliente

Conclusion

Conclusion

Our contribution:

- The adoption of hybrid approaches (such as MWE + clustering) brings advantages to the automatizing methods, in a way that the data does not need any previous human annotation to be used.
- MWEs can also bring domain knowledge that is implicit in the data.
- Using hybrid methods can help with future annotations.

Thank you for listening!

Acknowledgments

americanas sa

ufscar

Aix*Marseille université

Socialement engagée