An application of Spectral Clustering on Named Entity Recognition

Ali Josue Limon, Michele Cerú

New York University

ajl649@nyu.edu; mc3784@nyu.edu

December 7, 2016

Motivation

- Named Entity Recognition (NER) is one of the important parts of NLP. Neural Network methods for NER tasks in English performs well due to the vast amount of labeled data. However, given the lack of resources and labeled data in Spanish these models cannot be trained properly.

Goal and purpose

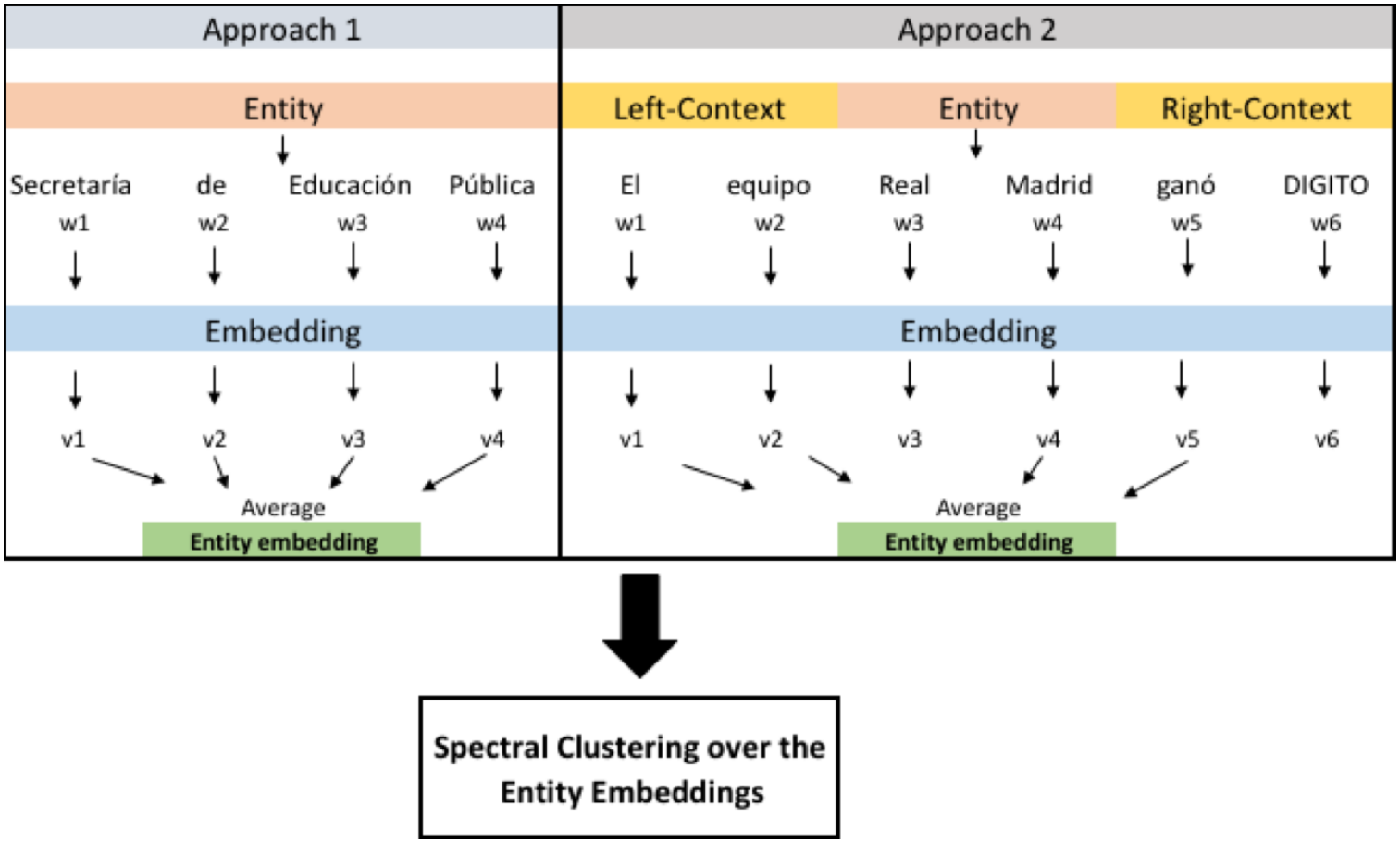
- Goal: Analyze the performance of Spectral Clustering combined with word2Vector embeddings in a Spanish Named Entity Recognition task
- Purpose: Develop new approaches that are suitable for a Named Entity Recognition task in languages with lack of labeled data

Named Entity Recognition

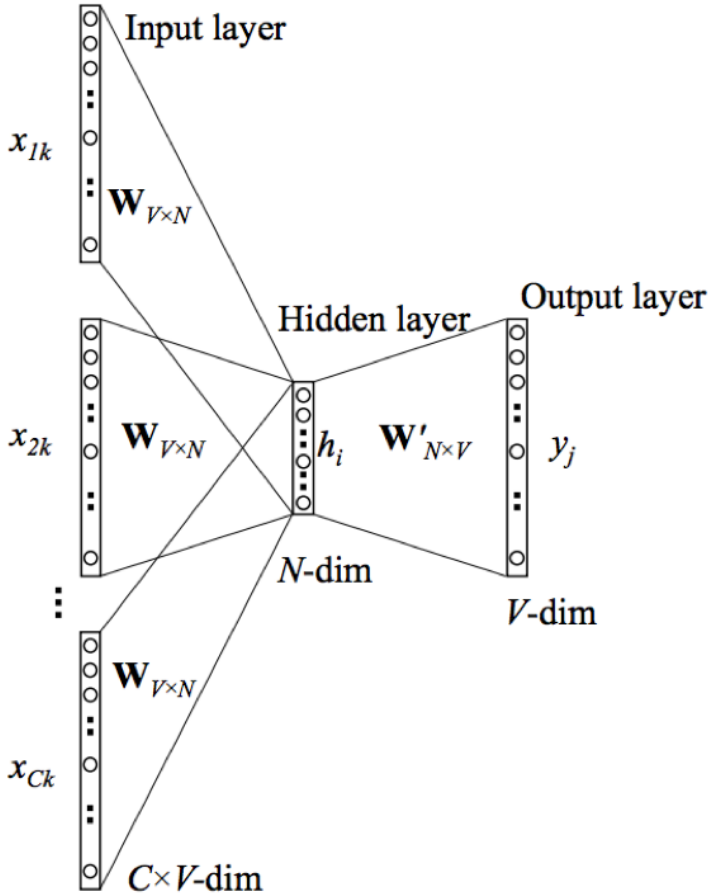
- It aims to find and classify expressions of special meaning in texts written in natural language. These expressions can belong to different predefined classes such as Person, Organization and Location.

  [PER Wolff] , currently a journalist in [LOC Argentina], played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

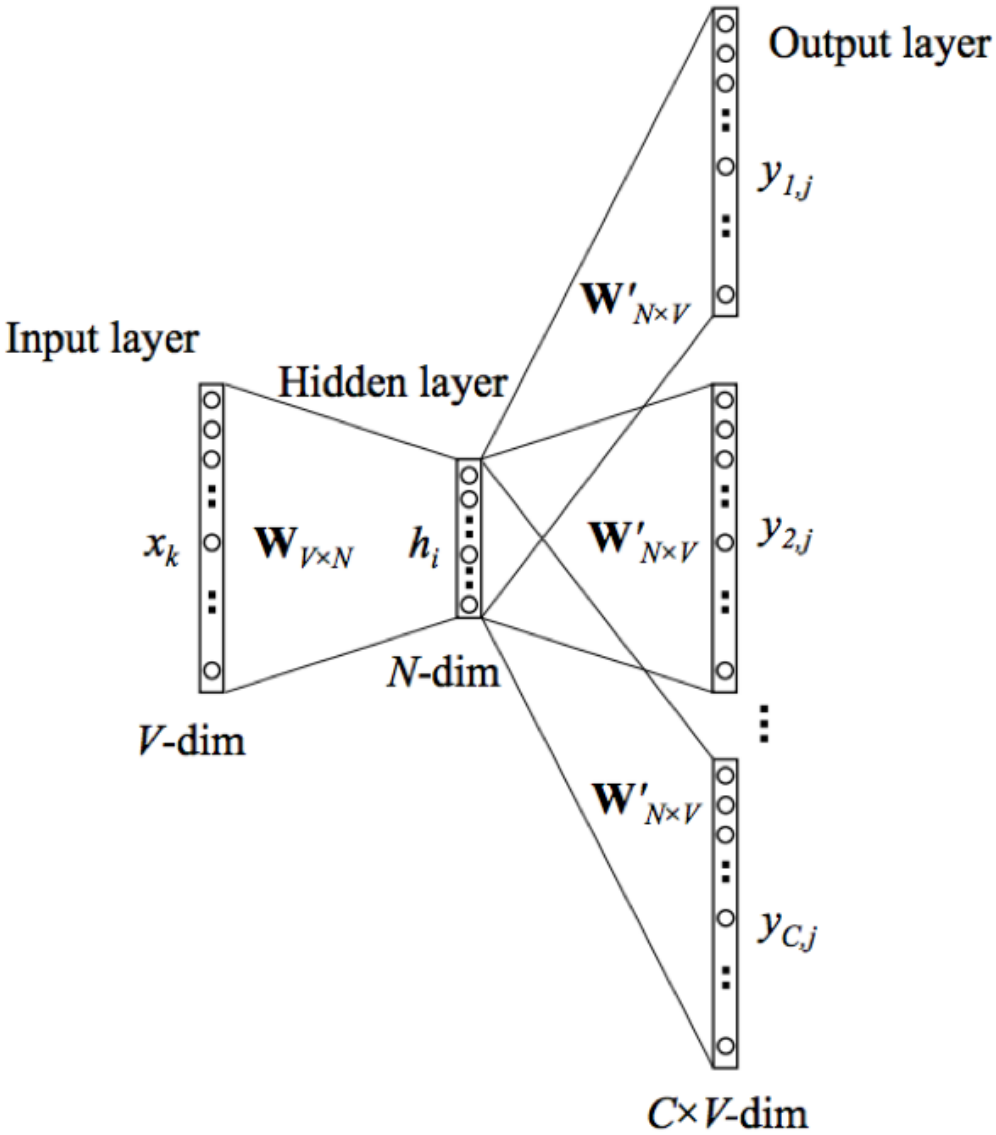
Problem Formulation



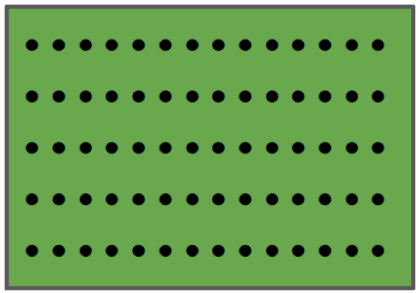
Word2Vector (CBOW)



Word2Vector (Skip-gram)



Dense word embedding



embedding matrix
E

one-hot vector
w_t

=

Dense word embedding
p_t

Dataset

- The *CoNLL 2002* Shared Task Spanish data. This dataset contains 11,752 sentences, 369,171 words and 20,451 Named Entities
- The Spanish Billion Words Corpus with a total of 1,420,665,810 raw words, 46,925,295 sentences and 3,817,833 unique tokens to train word embeddings

Model pipeline

- Train Word Embedding
- Average neighbour words in a window around the target
- Spectral clustering on the obtained vectors

Model parameters

	Parameters of the models	
WordEmbedding	CBOW	Embedding Size Context window = 5 Min Ocurrences = 5 Negative Sampling (Skip-gram)
	Skip-gram	
Spectral Clustering	Kernel	Cosine Similarity K-nearest neighbor $k(x, y) = exp(-\gamma \|\|x - y\|\|_1)$
Entity context		No context 1 word 2 words 3 words

Evaluation Criteria: Adjusted Rand index

In two clustering alternative C and C' there are 4 possible set of points:

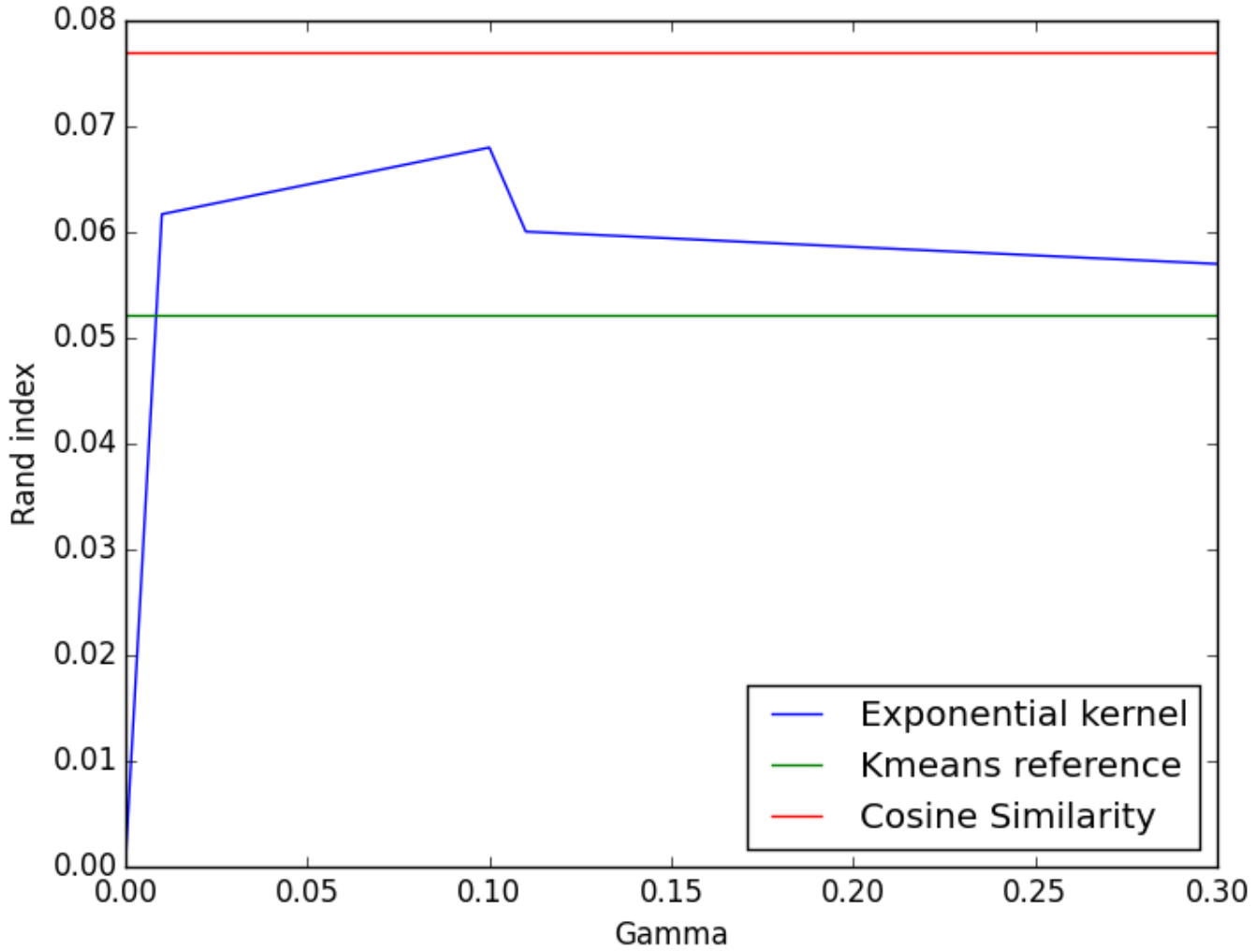
$$S_{11} = \{\text{pairs in the same cluster in C and C'}\}$$
$$S_{00} = \{\text{pairs in different cluster both in C and C'}\}$$
$$S_{10} = \{\text{pairs in the same cluste C but not in C'}\}$$
$$S_{01} = \{\text{pairs in the same cluste C' but not in C}\}$$

Rand index:

$$R = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} = \frac{n_{11} + n_{00}}{\binom{n}{2}}$$

Adjusted Rand index: Center the random clustering to 0

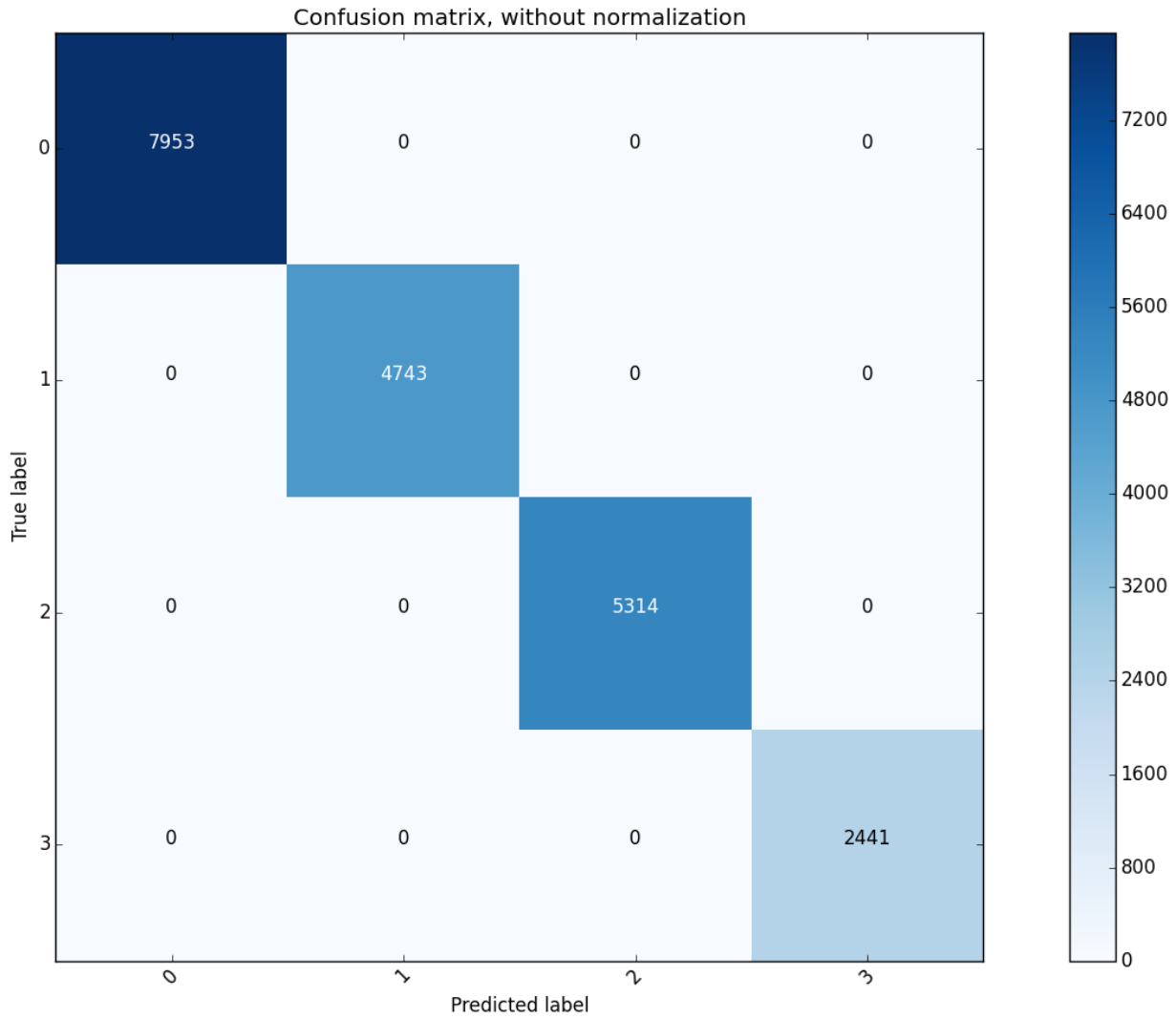
Results


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Rand index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Gamma
Exponential kernel
Kmeans reference
Cosine Similarity

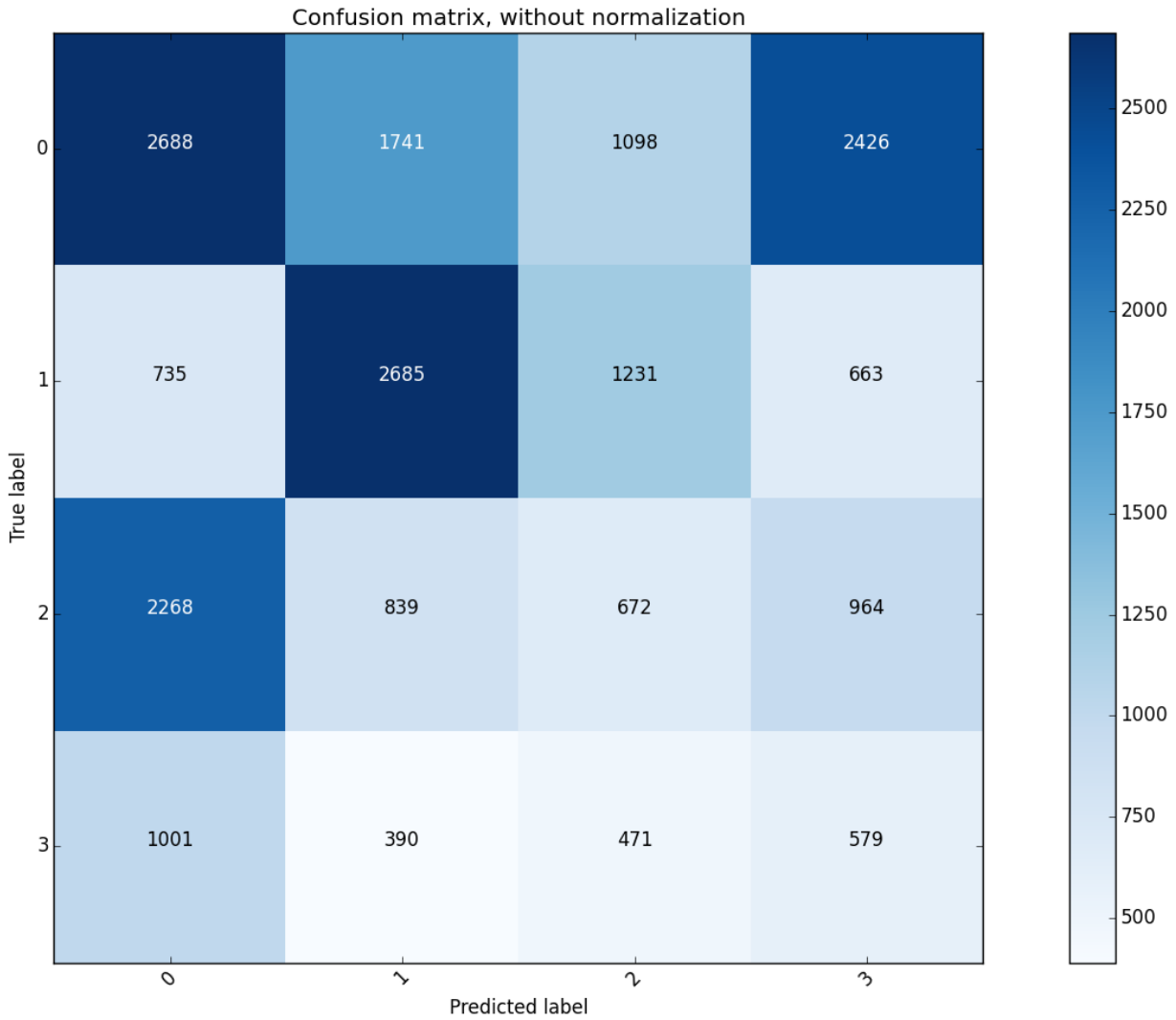
Data samples distribution ground truth

Cluster Label	Name Entity	Number of sample
0	Organization	7953
1	Location	4743
2	Person	5314
3	Miscellaneous	2441

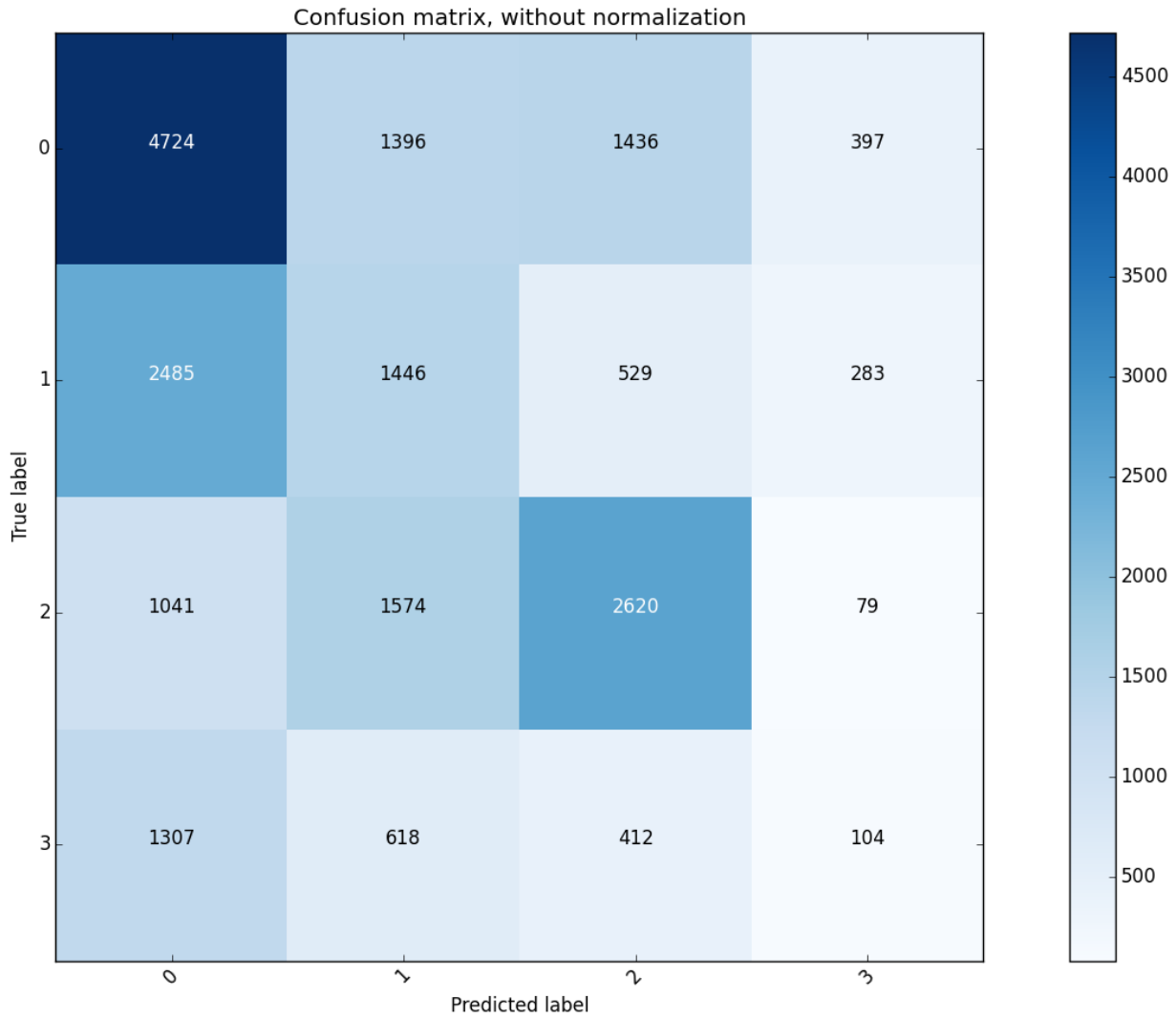
Confusion matrix: ideal clustering



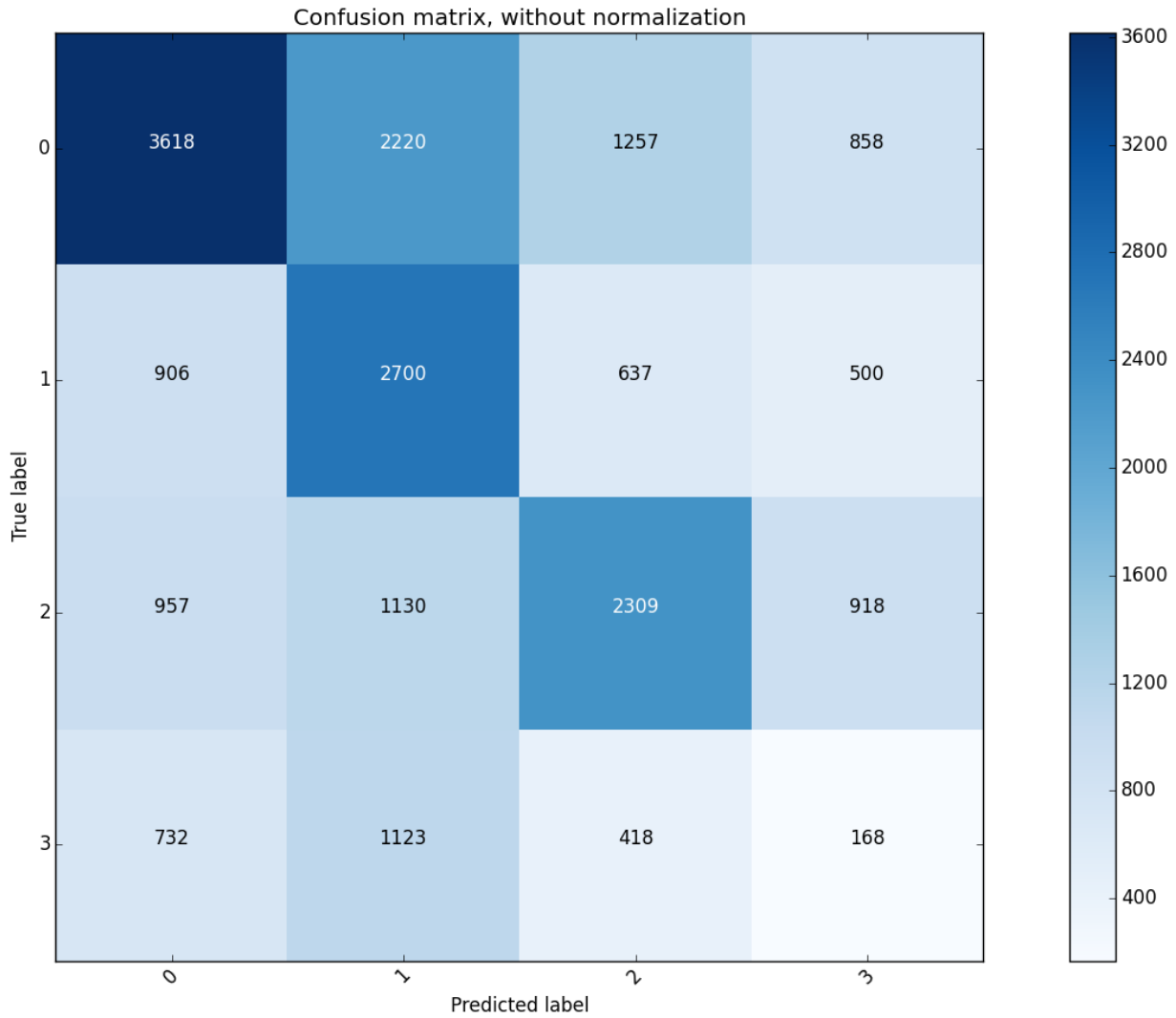
Confusion matrix: kmeans



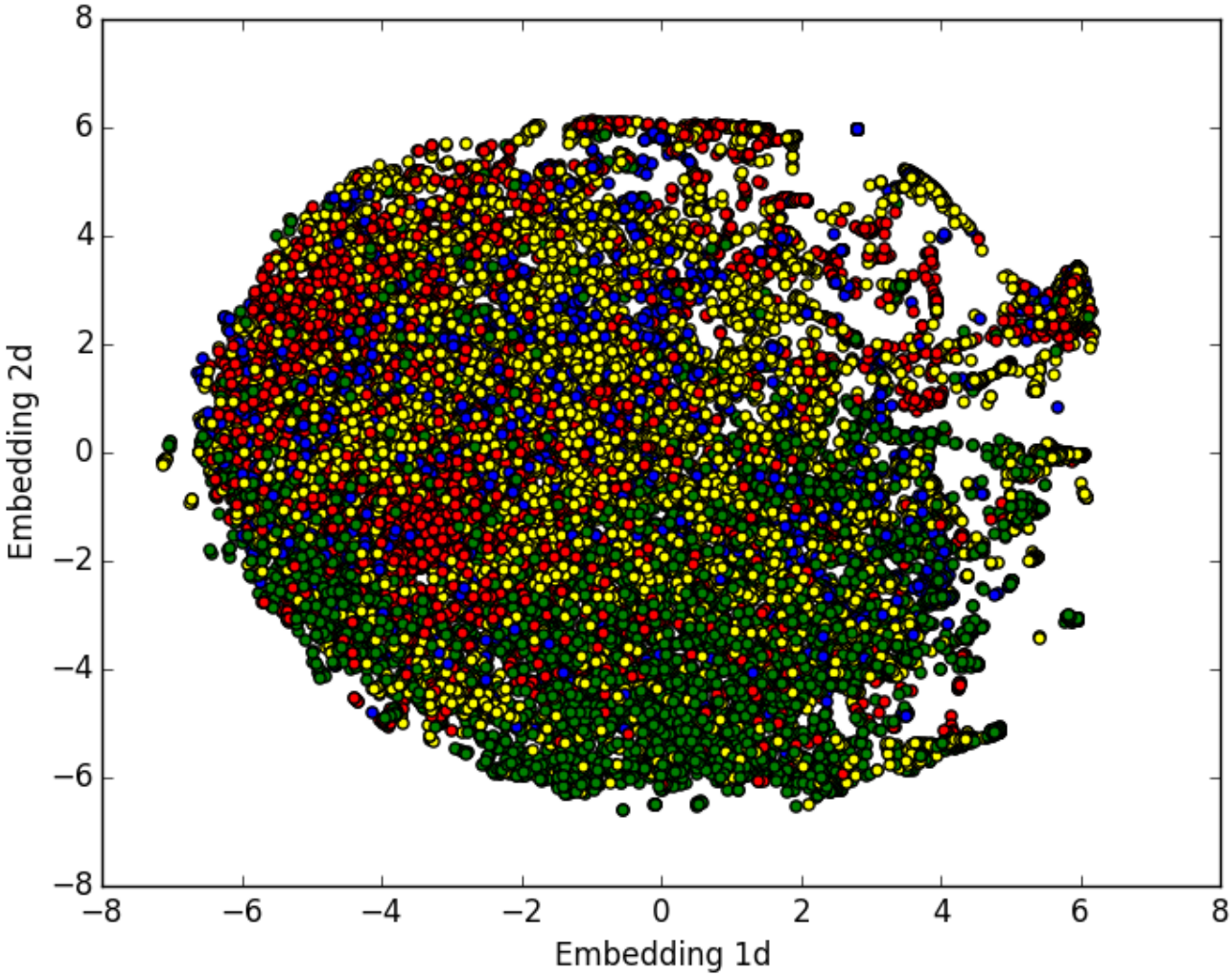
Confusion matrix: Cosine similarity



Confusion matrix: Exponential



Data visualisation in 2d



Conclusion

- Our approach does not seem to perform very well creating clusters according to the named entities labels

  Possible Reason: word embeddings representation encloses many different properties of the language, and named entity it is just one of them, consequently it is not possible to achieve a very good cut of the cluster