

Dealing with Polish language in NLP

Lessons learned at Grupa Pracuj

We will consider a multi-label task, e.g. automatically tagging a job offer

Data Scientist

Miejsce pracy: Warszawa / Nr ref.: GP/2018/4/AD

Jesteśmy jedną z najszybciej rozwijających się firm technologicznych w Europie Środkowej i wiodącym dostawcą rozwiązań internetowych wspomagających firmy m.in. w rekrutacji i budowaniu wizerunku pracodawcy. Należą do nas serwisy pracuj.pl i rabota.ua, a także inne biznesy związane z nowoczesnymi technologiami oraz rozwiązaniami „software as a service” (SAAS), np. platformy eRecruiter czy emplo. Zatrudniamy obecnie ponad 550 osób w Polsce i jesteśmy wyjątkowym miejscem pracy – już 8 razy otrzymaliśmy tytuł Najlepszego Pracodawcy.

Wykorzystujemy nowoczesne technologie informatyczne, zaawansowane algorytmy online, Business Intelligence oraz Big Data, aby dostarczyć klientom biznesowym najskuteczniejsze na rynku produkty online związane z rekrutacją pracowników.

Zakres obowiązków:

- projektowanie, testowanie, wdrażanie i utrzymanie systemów predykcyjnych na dużych zbiorach danych
- kolekcjonowanie oraz analiza danych z istniejących systemów danych do zadań machine learning
- wsparcie przy rozwoju produktów związanych z danymi
- dostarczanie analiz i raportów dotyczących zachowania użytkowników dla zespołów i osób decyzyjnych wewnątrz organizacji
- wsparcie przy utrzymaniu i rozwoju systemów analitycznych (hurtownie danych, systemy raportowe)

**Why does Polish language
need special treatment?**

Declension of a single Polish word – „stary” (old)

	r. m. l. poj.	r. ż. l. poj.	r. nij. l. poj	r. m. l. mn.	r. ż l. mn. .	r. nij. l. mn.
Mianownik	stary	stara	stare	starzy	stare	stare
Dopełniacz	starego	starej	starego	starych	starych	starych
Celownik	staremu	starej	staremu	starym	starym	starym
Biernik	starego	starą	stare	starych	stare	stare
Narzędnik	starym	starą	starym	starymi	starymi	starymi
Miejscownik	starym	starej	starym	starych	starych	starych
Wołacz	stary!	stara!	stare!	starzy!	stare!	stare!

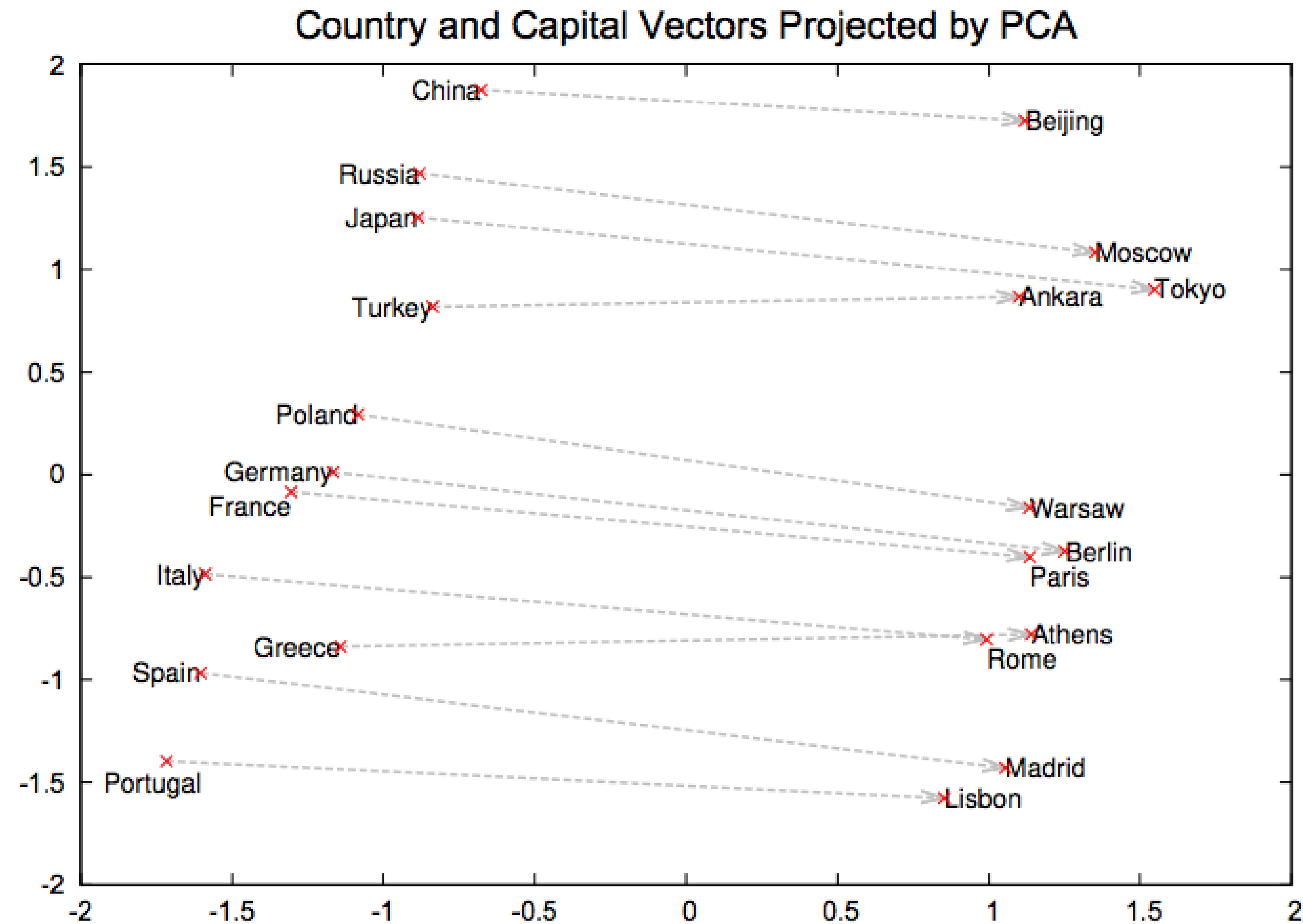
13 unique forms! (yes, I counted, but don't check pls :/)

Why is that a problem? Bag of Words example.

	sławy	stara	starej	starego	młody
observation 1	1	0	0	0	0
observation 2	0	0	0	0	1
observation 3	0	0	0	0	0

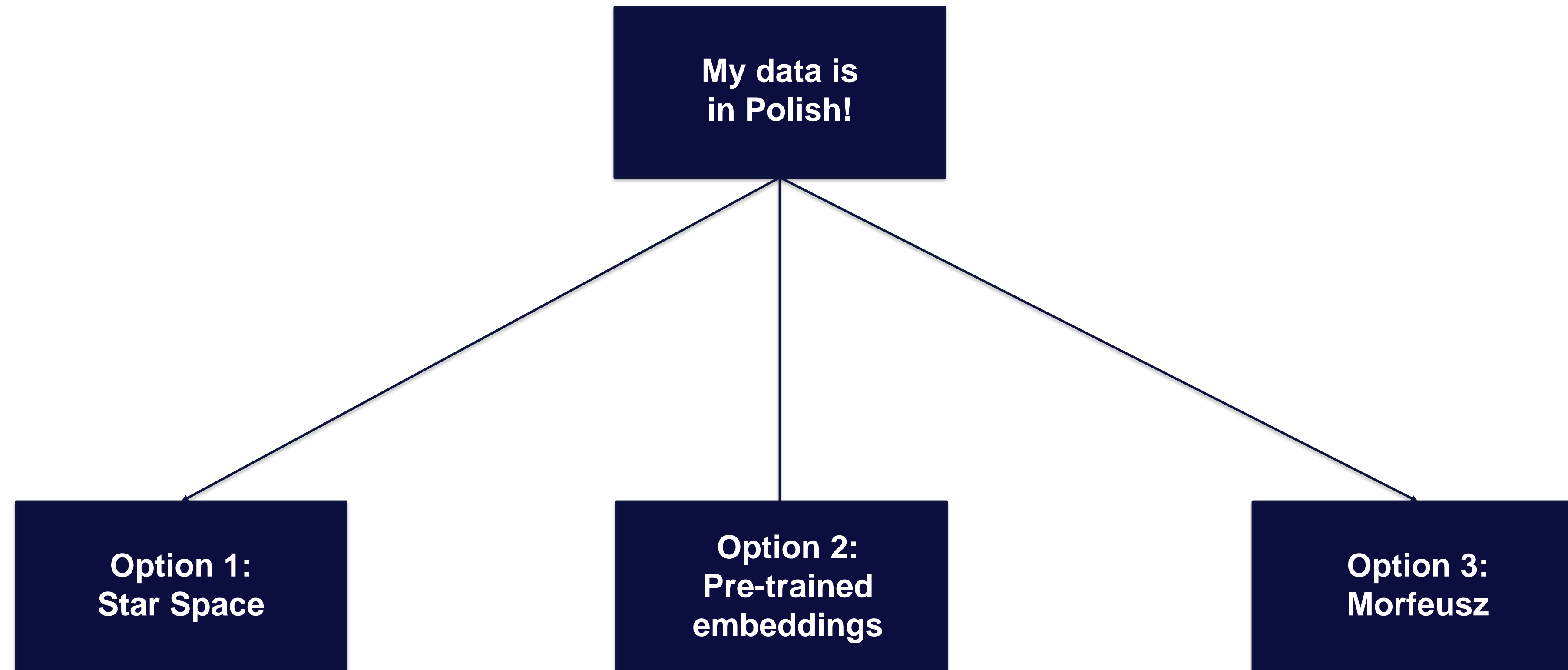
- This matrix will be huge (and sparse)!
- This might cause memory issues and requires a lot of data to be effective.


Continuous word embeddings deal with the first problem, but not the second.



Source: <https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>

Are we doomed? Naaah.



	Description	Pros	Cons
	"(Starspace) a general-purpose neural embedding model that can solve a wide variety of (NLP) problems"	<ul style="list-style-type: none">• Easy to use• Very fast development/training• Almost state-of-art results	<ul style="list-style-type: none">• Not easily customizable• Difficult to productionalize• "Almost" can make a difference ;)

Sources: <https://arxiv.org/abs/1709.03856> , <https://github.com/facebookresearch/StarSpace>

How does it work?

1. Generate positive entry pairs (usually your traditional learning data)
2. Generate negative entry pairs (e.g. by sampling from possible labels)
3. Calculate similarity
4. Loss function compares positive entries with negative ones.

Sources: <https://arxiv.org/abs/1709.03856> , <https://github.com/facebookresearch/StarSpace>

Description	Pros	Cons
Circumvent the problem of learning word embeddings by using other people's work :>	<ul style="list-style-type: none">• Easy to incorporate into own models• Robust - taught on a huge corpus (whole Wikipedia)	<ul style="list-style-type: none">• Can't easily deal with language mixes• Might have difficulties with jargon/ uncommon word meanings

Sources: <https://arxiv.org/pdf/1607.04606.pdf> , <https://github.com/facebookresearch/fastText/>

How does it work?

1. Download the word → embedding dictionary
2. Extract the word → token dictionary and the embedding matrix
3. Use tokens as input for the model
4. Use embedding matrix as weights for the embedding layer

Example in Keras:

```
# define the architecture
K = 5
no_filters = 1000
dict_length, vector_length = embedding_matrix.shape
no_predictors = lead_tags_tokenized.shape[1]

sequence_input = Input(shape=(MAX_SEQUENCE_LENGTH,), dtype='int32')
embedded_sequences = Embedding(dict_length, vector_length,
                               weights=[embedding_matrix],
                               input_length=vector_length, trainable=False)(sequence_input)

x = Conv1D(no_filters, K, activation='tanh', padding='same', input_shape=(300, 300))(embedded_sequences)
x = MaxPooling1D(vector_length)(x)
x = Activation('tanh')(x)
x = Flatten()(x)

preds = Dense(no_predictors, activation='sigmoid', kernel_initializer='uniform')(x)

model = Model(sequence_input, preds)
model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=[])
```

Sources: <https://github.com/facebookresearch/fastText/>, <https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html>

Description	Pros	Cons
<p>Lubię grać w piłkę</p> <p>↓</p> <p>Lubić grać w piłka</p>	<ul style="list-style-type: none">• Can be used to build BOW models• Allows to learn embeddings with a smaller corpus• Can be used with Python2, Java, C++	<ul style="list-style-type: none">• Very difficult to work with• No Python 3 ☹️• Performance issues

How does it work?

1. Try to install morfeusz on your computer
2. Cry, because it's difficult to make it work
3. Finally succeed after struggling for a bit
4. Convert your texts into lemmatized version
5. Proceed as in Option 2 (except don't provide embedding matrix, because you won't have any!)

Sources: <http://sgjp.pl/morfeusz/>

What should I use then?

Option 1: Star Space

- Fairly standard problem
- Not that concerned with extra 1-2 p.p. of success metric
- Different languages mixed

Option 2: Pre-trained embeddings

- Not enough data to learn embeddings
- Standard vocabulary

Option 3: Morfeusz

- Not enough data to learn embeddings
- Very particular vocabulary and/or its use (context)
- Dislike deep learning

We are hiring !

Thank you for your time. Any questions? 😊

jan.zysko@gmail.com