Practical No. 1

Mateusz Kwiatkowski
mk457176@students.mimuw.edu.pl

March 14, 2023

1 Word Vectors

In this exercise I got two co-occurrence plots for polish words.

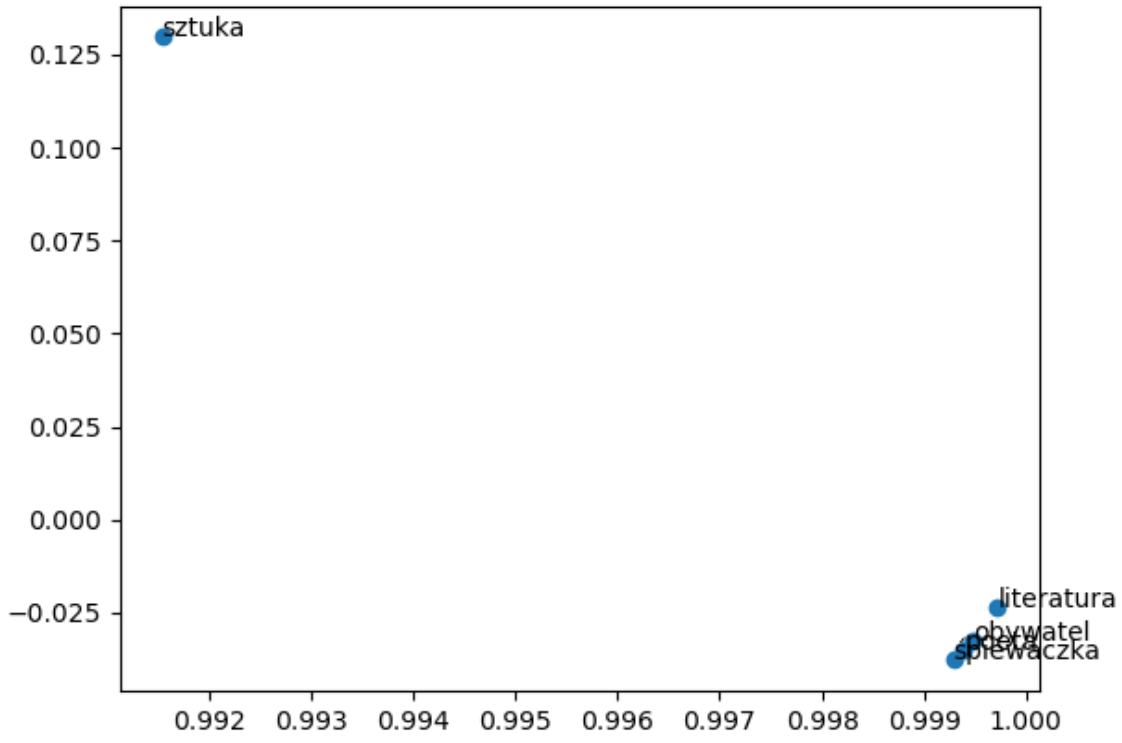

Figure 1: Normalized plot

As you can see, words "poeta", "obywatel" and "śpiewaczka" cluster together, which makes sense, because they all describe people, so they can be used in the similar context.

However, it's strange that word "literatura" is close to them. It seems that it should be closer to "sztuka".

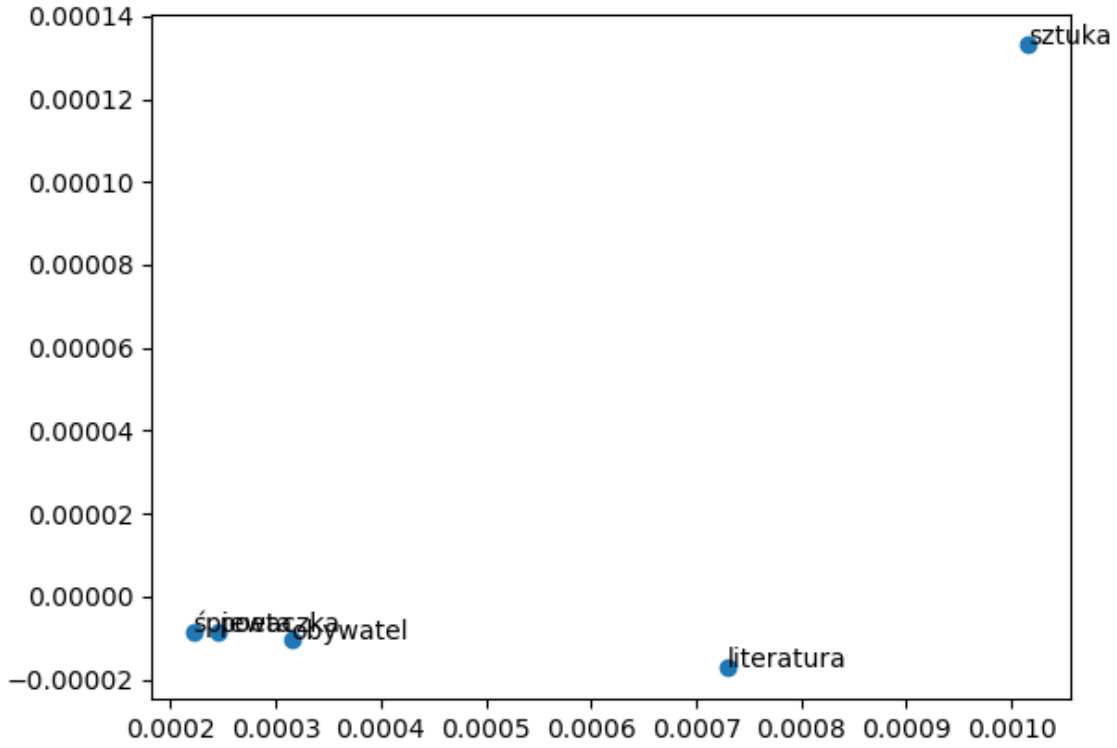

Figure 2: Unnormalized plot

Unnormalized plot looks quite similar, because again words "poeta", "obywatel" and "śpiewaczka" cluster. The biggest difference is that "literatura" lies away from them. This more reasonable result.

2 Prediction-Based Word Vectors

a)

In this task I again made a diagram to see how the words are arranged in a two-dimensional space. It looks a bit different than plots generated earlier from co-occurrance matrix. In this case, the words related to people are far apart. Close to each other are the words "art" and "artistic" which makes a lot of sense.

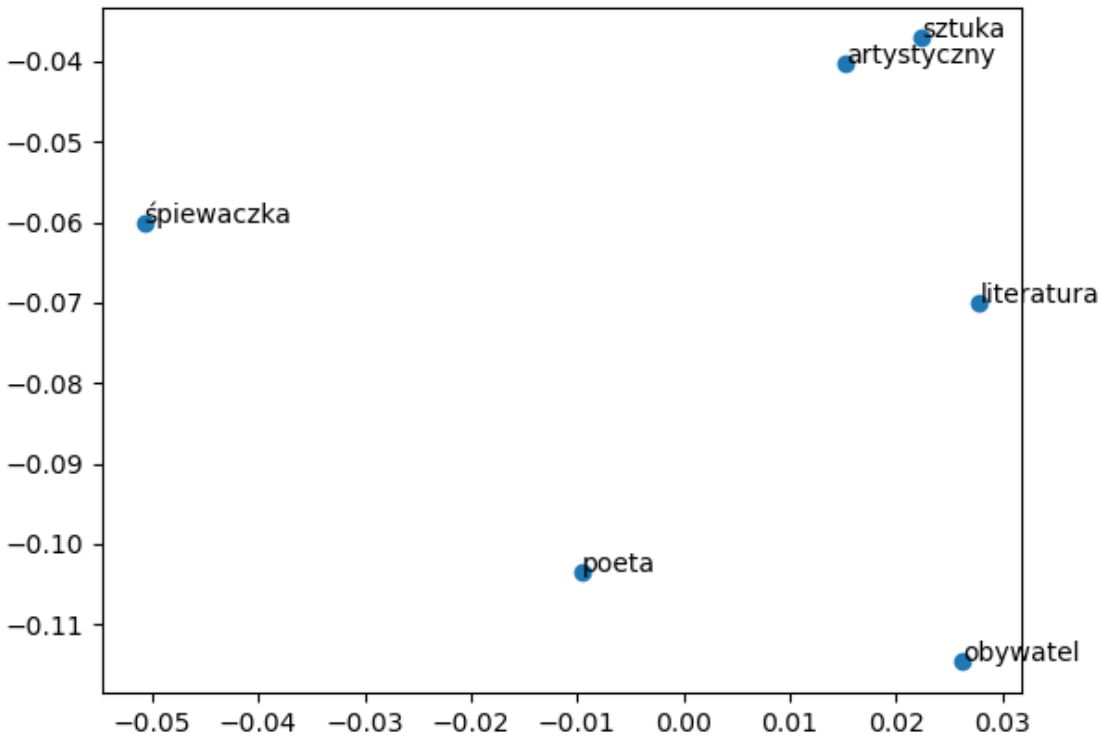

Figure 3: Unnormalized plot

b)

In this subsection, I needed to find a word that appears in several meanings. I chose the word 'blok'. Among the most similar words both 'barak' and 'budynek' appear, as well as 'sześcian'. The few words I tried to substitute did not work. I think this is due to the fact that one meaning was far more common than the other and dominated the results.

c)

I found three words: ("dobry", "poprawny", "zły"). It turns out that "dobry" and "zły" are closer than "dobry" and "poprawny". In my opinion, this happened because we can use the words "dobry" and "zły" in exactly the same sentences, for example: "To był dobry dzień"/"To był zły dzień". Nobody would say "To był poprawny dzień". "Dobry" i "poprawny" have very similar meaning, but they aren't used in the same context.

d)

I found the analogy dziadek:ojciec :: babcia:matka. I think that this analogy is obvious and the algorithm gave correct answer on the first place.

e)

I wanted to find some more complex analogy and I came up with the idea: serce:przedsionek :: dom:przedpokój. Unfortunately, the algorith didn't find it and gave answer: serce:przedsionek :: dom:suterena, which is a little counter-intuitive.

f, g and h)

In this subsection I needed to search for bias in word vectors. For the enquiry mężczyzna:szef :: kobieta:? I get answers "własika" and "antyterrorystyczny", which is a total absurd. I also get analogy kucharz:programista :: kucharka:implementacja. This shows the weaknesses of this algorithm. Probably this mistakes appears because the set of texts was too small.

i)

I repeated the same tasks in English. The example of polysemous word is "left", becouse it can be used in two meanings, for example "left side" and "left from the house". I repeated the experiment for words ("good", "correct", "bad") and got the same result. I also tried to find the same analogies and results were: grandfather:father :: grandmother:mother and heart:auricle :: house:houses (incorrect). The example of bias is: woman:boss :: man:supremo.