NLP

- SCIA 2023 -

Jules DORBEAU - Noé JENN-TREYER - Yacine ANANE - Adrien HOUPERT

Theoretical questions (9 points)

1 (2 points) Explain with your own words, using a short paragraph for each, what are:

1.1 Phonetics and phonology

Phonetics is a branch of linguistics that studies the physical properties (articulatory, acoustic,...) of sounds in spoken communication. It is interested in the sounds themselves, independently of their functioning with each other. And it allows us to describe precisely the way words are pronounced.

Unlike phonetics, phonology is specific to a given language. It is another branch of linguistics that studies sounds from a functional point of view, it studies the way sounds of a language are arranged to form sentences.

1.2 Morphology and syntax

Morphology is a branch of linguistics that involves the study of the grammatical structure of words and how words are formed for a given language. Morphology studies the relationship between morphemes, the smallest meaningful unit of a word, and how these units can be arranged to create new words or new forms of the same word.

Syntax is the branch of linguistics that studies the way words combine to form sentences or statements in a language. It, therefore, studies the order of words in a sentence, grammatical categories, and grammatical functions.

1.3 Semantics and pragmatics

Semantics is the study of the meaning of words, phrases, and sentences. In semantic analysis, there is always an attempt to focus on what the words conventionally mean, rather than on what an individual speaker might want them to mean on a particular occasion.

As opposed to semantics, pragmatics considers the context of sentences and studies how context contributes to meaning. It allows a better evaluation of the utilization of language in human-spoken communication.

2 (1 point) What is the difference between stemming and lemmatization?

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

2.1 How do they both work?

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.

Lemmatization, on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

2.2 What are the pros and cons of both methods?

Stemming is fast because it requires little computing power, but its results are sometimes bad.

The main advantage of lemmatization is that it is more accurate than stemming. For instance, a good lemmatizer would know that the root of better is good but a stemmer cannot find the root of this word. However, lemmatization involves deriving the meaning of a word from something like a dictionary, it's very time consuming. So most lemmatization algorithms are slower compared to their stemming counterparts.

3 (1 point) On logistic regression:

3.1 How does stochastic gradient descent work?

Gradient descent is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function. In order to do that, we need to follow these steps:

- 1) Compute the gradient of the function.

- 2) Pick a random initial value for the parameters.
- 3) Update the gradient function by plugging in the parameter values.
- 4) Calculate the step sizes for each feature as: step size = gradient * learning rate.
- 5) Calculate the new parameters as : new params = old params -step size
- 6) Repeat steps 3 to 5 until the gradient is almost 0.

The problem with this approach is that it can be slow to compute the gradient because there might be a lot of training data involved in the error. The stochastic gradient descent is a solution to this problem because we will compute the error using a random data point for each iteration instead of doing it for all data at each iteration.

3.2 What is the role of the learning rate?

The "learning rate" is a parameter which heavily influences the convergence of the algorithm. If the learning rate has a high value, the algorithm takes huge steps down the slope and it might miss the minimum point by jumping over it.

3.3 Will it always find the global minimum?

Although the randomness of the SGD can sometimes allow it to escape from a local minimum to find a better one, the algorithm can still get stuck in local minimums.

4 (1 point) What problems does TF-iDF try to solve?

4.1 What is the TF part for?

TF stands for "Term frequency", it's the number of times a term occurs in a document.

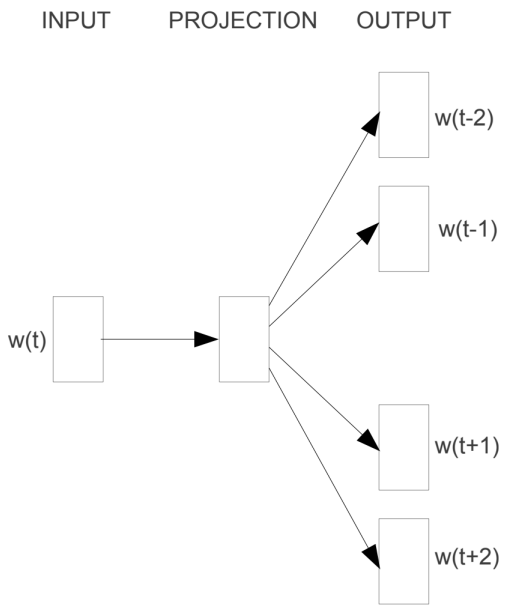
4.2 What is the iDF part for?

iDF stands for "inverse document frequency". iDF is used to decrease the weight of terms that appear very frequently in the document set and increase the weight of terms that appear rarely.

5 (2 point) Summarize how the skip-gram method of Word2Vec works using a couple of paragraphs.

5.1 How does it uses the fact that two words appearing in similar contexts are likely to have similar meanings?

The skip-gram of Word2Vec is a method used for computing word embeddings. More precisely, it is an architecture that uses the central word to predict the surrounding words (instead of the CBow of Word2Vec that uses the surrounding words to predict the center word).The skip-gram objective function sums the log probabilities of the surrounding words to the left and right of the target word.



Skip-gram

This is why, two words appearing in similar contexts are likely to have similar meanings. As the surrounding words are generated depending of the center word, two words with similar meanings are likely to be used in similar contexts.

6 (1 point) What are the differences between an RNN and an LSTM?

The LSTM is a form of RNN but adding memory cells and gates ("input", "output" and "forget") that can be opened or closed. The "input" gate is used to update de cells status by giving them a score of importance, between 0 and 1, respectively important and not important. The output gate have the information of previous inputs and calculates the next hidden state value. The forget gate can decide which information needs attention or needs to be ignored.

6.1 What problem is an LSTM trying to solve compared to a basic RNN?

The LSTM is trying to solve the vanishing gradient problem, which means the gradient of the loss function gradually lose information over time, by adding a long term memory system with memory cells and the gates system which allows a better control over the information and training.

7 (1 point) What would you expect if we use one of our classifiers trained on IMDB on Twitter data, and why?

If the model trained on IMDB is used on English tweets, we expect the model to have bad results for multiple reasons:

- Tweets are often neither negative nor positive.
- Tweets have a limitation on the number of characters causing users to use abbreviations that the model might not know.
- Because of the limit of characters, the model has less information to work with compared to IMDB reviews.
- There is a lot of memes, internet reference, and internet culture used to communicate on Twitter. These words are less likely to appear on IMDB so the model might not understand these terms.
- IMDB and Twitter may not have the same user profile (average age, hobbies, etc) which could cause differences in the way people express themselves on the 2 platforms.