University of Amsterdam
Economics & Business

Data Science Minor - 2020/2021 - Sem. 1, Period 3

Text Retrieval Text Mining
BERT and Course Recap

Julien ROSSI

Week 4

After this lecture, you will:

- Know what is BERT

Previously in Text Representations

N-Gram Language Model

- Conditional Probabilities of a word occurring, knowing N-1 previous words

Word2Vec Skip-Gram

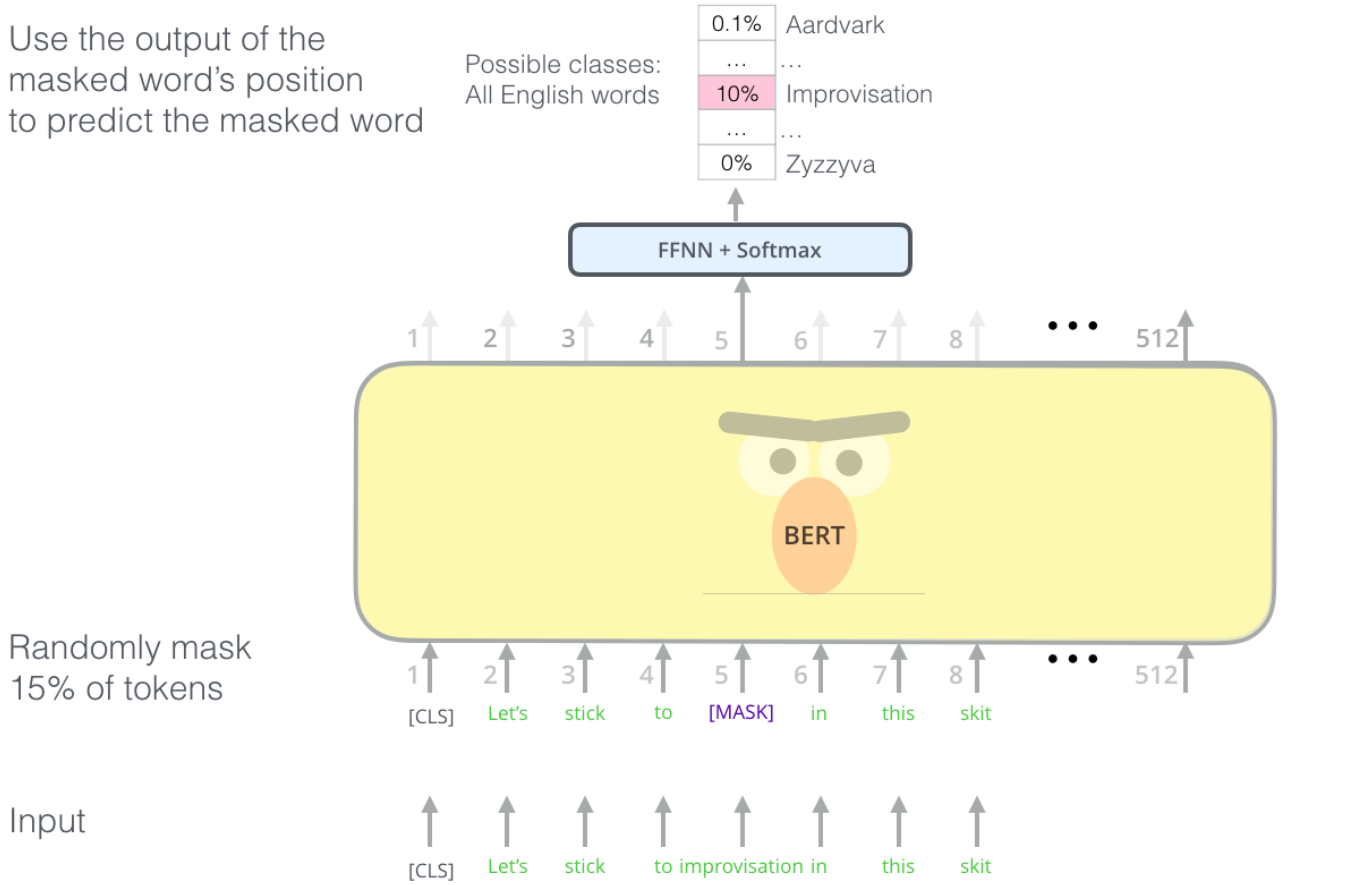
- Generate probability distribution of context words given a central word with a bottleneck neural network
- Use network weights as word vectors

BERT (not for exam)

Devlin & al., 2018
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

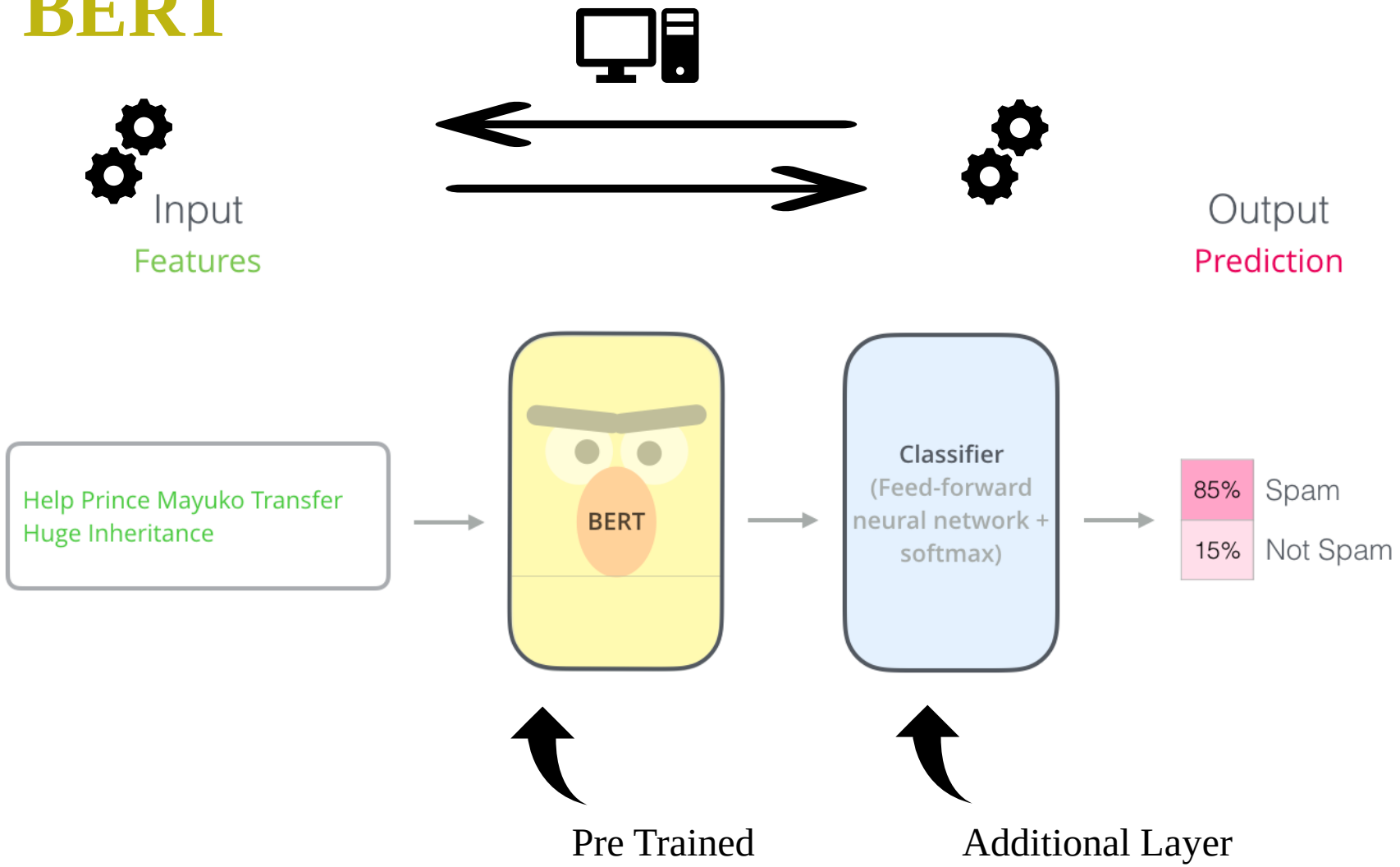
Inside BERT:
12-24 Transformer Encoder Layers
(Vaswani & al., 2017, Attention is All you Need)

Language Modeling:
Learn how to predict a missing word from a sentence.



Images from codemotion.com http://jalammar.github.io/illustrated-bert/

BERT


Input
Features
Output
Prediction
Help Prince Mayuko Transfer Huge Inheritance
BERT
Classifier
(Feed-forward neural network + softmax)
85%
Spam
15%
Not Spam
Pre Trained
Additional Layer

BERT

- Google made Pretrained Models available (~7k$ computation costs)
- Trained on Google News / Wikipedia / Books
- https://github.com/google-research/bert
- https://github.com/huggingface/transformers

After BERT

Following the same 2-phase principle:

- Pre-Training of a very large neural network: **expensive**
- Fine-Tuning on a task: **cheap**

- Using Pre-Trained models, as pre-training is more and more expensive

End of Course

This was the last content for this course.

Thanks for your attendance and participation.

Please fill in the Course Evaluation

Exam

Content:

- 6 Multiple Choice Questions (6 * 5 = 30pts)
- 6 True/False Questions (6 * 5 = 30pts)
- 1 Essay Question (1 * 40 = 40pts)

Canvas Quiz:

- Available 09:00
- Deadline 11:00
- Can be taken **only once**
- Each question can be visited **only once**

Exam

Multiple Choice Question

"You study a corpus of 1 million books, for a total of 2 billion tokens. The initial vocabulary size is 200,000. 40,000 terms appear in less than 5 books. You keep in vocabulary all terms with a minimum document frequency of 5.

1. The vocabulary size is now 160,000
2. The vocabulary size is 240,000
3. The vocabulary size is 2 billion + 40,000
4. None of the answers are correct.

Exam

True / False

"You study a corpus of 1 million books, for a total of 2 billion tokens. The initial vocabulary size is 200,000. LDA is used for Topic Modeling with K=200 topics.

Each document is represented with a vector with 1,000 dimensions.

1. True
2. False

Exam

Essay

"You study the contracts of your company, trying to automatically differentiate lease contracts from work contracts. In addition, you want to identify 6 different types of lease contracts, based on how they are built.

Describe how you would tackle the issue.

- 1-liner will yield only a few points ("Bag of Words with TF-DF")
- Give motivation, make hypothesis, …
  - What do you want to capture?
  - Is it visible at lexical level, or semantic level?
  - Is it about topics, or about meaning?

This course in a nutshell

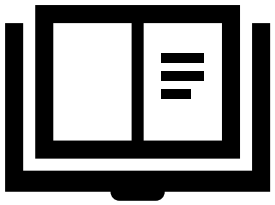
Represent texts in a way that a computer can assist us to perform useful tasks

CLASSIFICATION
What kind of book is this?

RETRIEVAL
Which books are about dinosaurs?

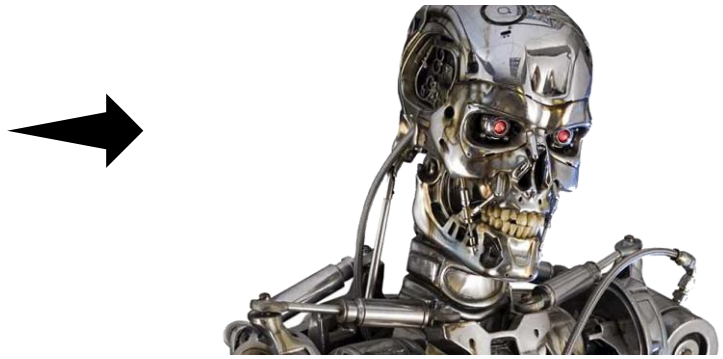
CLUSTERING
To what group of books is it the closest?

This course in a nutshell

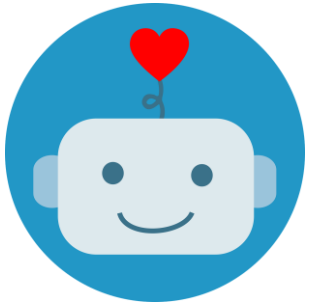


In “Small Gods”, Terry Pratchett

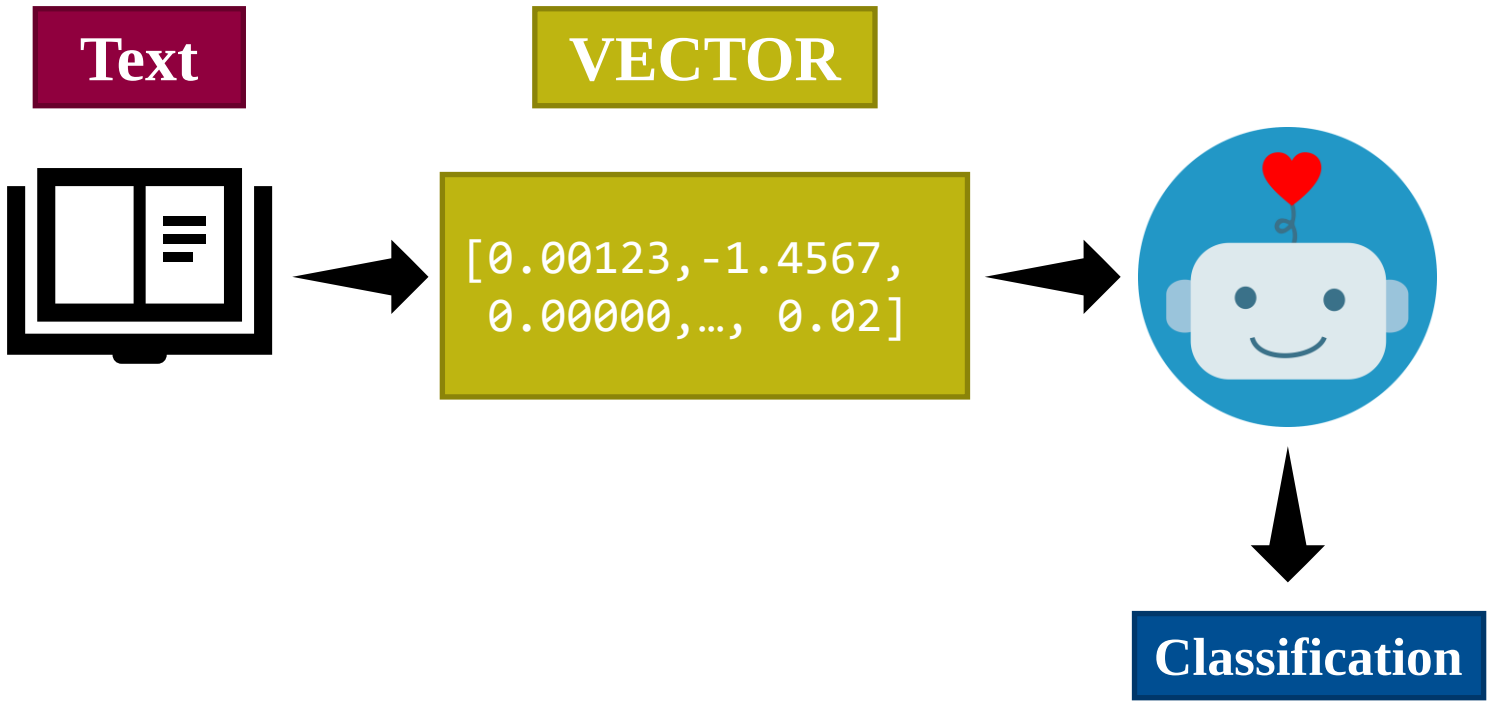
“Time is a drug. Too much of it kills you.”



```
[0.00123,-1.4567,
 0.00000,…, 0.02]
```



This course in a nutshell



This course in a nutshell

LEXICAL REPRESENTATIONS:

- Bag of Words, TF-IDF
- Vector dimensions = terms in vocabulary
- Pre-Processing (Stemming, Stopping, Lemmatizing, …)

- Strength: leverage important words, choice of words / colocations
- Weakness: difficulties to deal with paraphrasing, synonyms

This course in a nutshell

SEMANTIC REPRESENTATIONS

- Topic Modeling
- SVD, LDA
- Vector dimensions = Topics

- Strength: describe a document as which subjects it deals with
- Weakness: writing style

This course in a nutshell

N-GRAM Language Model

- Build conditional probabilities from corpus
- Naïve classification

- Strength: writing style
- Weakness: dealing with paraphrasing, synonyms

This course in a nutshell

Word2Vec

- 1 word = 1 vector
- Learned from corpus

- Strength: synonymy, paraphrasing
- Weakness: writing style