Natural Language

Practical Classes

Luísa Coheur

2025

P5

Text as **Sparse Vectors** and Feature-based **Machine Learning**

Image generated by ChatGPT

- **Summary**:
  - Text as (sparse) vectors
  - Feature based machine learning
- **Operational objectives**:
  - Practice the representation of text as sparse vectors
  - Understand feature based machine learning via a case study
- **This class needs**: paper, a pen/pencil and a computer
- **Class material**: these guidelines and a jupyter notebook

Another client!

You have a new email:

Dear NLP Detective,

I hope this message finds you well. I am writing to you about a project I have been working on with a friend. We have been organizing Dungeons & Dragons (D&D) games together. My friend is very systematic and wants to categorize the spells, which is quite a labor-intensive task. I have been hearing a lot about AI and NLP lately. I was wondering if there is any technique or tool that could help automate the categorization of these spells. If such a solution exists, would you be interested in developing a program to assist me in classifying the spells?
Thank you very much for your consideration.
Warm regards,
Sam

“Spells? That sounds funny![1]” You recall what you have learned in your recent NL classes about text as vectors and feature engineering. This will be an excellent scenario to test and practice your acquired knowledge.
You respond to Sam:

Yes, there is such a solution, and I will be glad to accept your challenge.

While you wait for a response, you decide to do some exercises.

1 Towards expertise

1.1 Language as vectors

You read an amazing book about zombies. Today you decide that you want to read a similar book. You run to the nearest library and search for similar books by typing the title of your book:

“The happy zombie in your room” (d0 – let’s call it “document 0”).

The system returns:

d1: Happy rooms
d2: The zombie room
d3: The flowers in your room

Now the question is: which book should you read? You decide to follow the distributional approach that views language as vectors.

1. You represent the books’ titles as vectors. Lowercase the text, transform plurals into singulars and remove stop-words (the, in, your). Considering a raw count, create the

[1] Not necessarily the word; the concept.

corresponding table (columns d0-d3 (documents 0-3), lines “happy”, “zombie”, “room” and “flower”).

2. By using the cosine similarity, find the similarity between d0 and each one of the other titles. Which book would you choose with that technique?

3. Now, let D be the whole set of titles/documents d0, d1, d2 and d3. Find tf-idf(zombie, d0, D) and tf-idf(room, d0, D) (use the formula studied in class, but the absolute frequency for the tf):

4. Make a short comment on the difference between the value of tf-idf you have obtained and the word “room” raw count.

1.2 Feature-based sequence labelling

When you have just finished the exercises, Sam replies to your email:

Dear NLP Detective:

Thank you for accepting this job. In attach you have the dataset “cast”, a modified version of the dataset dnd-5e-spells[2]*. The ideia is to have a program that receives a spell and returns a word (column “category”) from file “cast”. Let me know the results achieved.*
Warm regards,
Sam

You spend some time looking at the corpus. Then, you use the given jupyter notebook.

1. You run the jupyter notebook. Which is the best model?
2. Try to improve results. Play with: a) The size of the vectors; b) Data (in)balance; c) Pre-processing; d) Extra features; e) Other

Not an easy task. You send an email to Sam, reporting the miserable results you achieved. Sam replies shortly after:

Dear NLP Detective,

Very disappointing. As you can understand, I will not pay for a program that works so poorly. I hope you will be able to improve it.
Warm regards,
Sam

“Oh, shoot.” – you think – “Well, tomorrow is a new day, and maybe I will have a new client” . Anyway, now, you have to study for the NL MAP1. No stress. Just another test.

[2] `https://github.com/TheDataRogue/dnd-5e-spells`