Natural Language
Practical Classes

Luísa Coheur and Rui Henriques

2025

P2

Dealing with corpora and falling in love with statistics

Image generated by ChatGPT

- **Summary**:
  - Data annotation: guidelines and agreements
  - Enriching and understanding data: taking advantage of statistics
- **Operational objectives**:
  - Make students aware of the importance of having good guidelines before the annotation process
  - Provide students with the experience of annotating a non-trivial dataset
  - Provide students with a didactic and practical overview of statistics that can be used to enrich and better understand a dataset
- **This class needs**: a computer
- **Class material**: datasets for annotation and notebooks.

You had a strange dream. Two people dressed as detectives were discussing statistics. One was called "Natural Language (NL)" and the other "Processing (P)". When you woke up, you could still remember their dialogue:

NL: How do we guarantee that our answers are statistically significant?

P: We use statistical hypothesis testing.

NL: Can you recall me the basics?

P: You define the null hypothesis (H_0) for no association/relationship/difference.

NL: Go on...

P: Choose significance level (α), the probability threshold for rejecting H_0 (often 0.05).

NL: Got it. Go on...

P: Select test statistic (depends on data type and assumptions) and then take a decision.

NL: A decision?

P: Yes: compute the test statistic and associated p-value (p), then:

- If $p \leq \alpha \rightarrow$ reject H_0
- If $p > \alpha \rightarrow$ fail to reject H_0

NL: Thanks, man, you are a real NLP detective!

You wake up a little tired from the strange dream. You hadn't even remembered that you knew so much about statistics. Anyway, your decision is taken: you will start your own company, the NLP Detective. You announce it in LinkedIn and tell all your family and friends about it.

While you have no clients, you decide to sign up (without actually needing to enrol) for Amazon Mechanical Turk[1], a crowdsourcing marketplace that enables "Requesters" to engage "Workers" to perform various tasks. You also enrol in other similar sites in which people are paid to annotate data. After a while, you receive your first request from a small startup called _____ ← create a nice name for the startup. With it, you receive a disclaimer:

This dataset contains material that some readers may find offensive. This includes depictions of violence, and strong language.

1 Your first task: annotate the jokes dataset

You receive an XLSX dataset named Jokes[2]. You give a brief overview: 150 jokes and 6 labels (Stereotype, Error, Repetition/Theme, Attack/Offense, Misunderstanding, and Wordplay). But what exactly do they mean? Some labels are more or less obvious, but others... You complain, and eventually, they send you the guidelines (provided in appendix[3]).

You start the annotation. You contact a colleague who is also annotating the same jokes to ensure that you're doing it correctly. Your tasks:

1. Annotate 5 jokes, namely jokes 6, 28, 49, 62 and 73.

[1] `https://www.mturk.com`
[2] Thanks to Sebastião Caldas for this dataset!
[3] Thanks to Sebastião Caldas for these guidelines!

2. Contrast your annotations with the ones from your colleague, computing the agreement for each criterion. Which criteria were harder to agree on?
   You find a notebook, P2_agreements.ipynb, that can help. You discover that you can calculate a p_value with it. Nice. Your dream was amazingly useful!
3. You noticed that you could contrast your answers with a third colleague. You find yourself no longer able to apply the Cohen's Kappa metric. What should you do?

You decide to submit your curated annotations to the startup. After a while, you receive an email saying that the annotations are reasonable – you have passed the first stage – and to become a paid annotator, you need to pass stage 2, which involves annotating outdated words in the 32 volumes of the Encyclopaedia Britannica. Oh, dear!

2 Statistics? I Never thought I could master them!

It's time to come back down to earth and work on the exercises your NL professors have prepared for you. Today's class will be about statistics. You remember that, during a theoretical class, your professor explained you would have an expert in statistics this year among us, and that she would take advantage of his knowledge to learn how to do a number of things in NL using statistics. The target will be an Amazon dataset of fragrance reviews (P2_dataset_reviews.csv). The plan is as follows:

1. Enrich the dataset (not just statistics);
2. Study the dataset using basic statistics, while also trying to understand (empirical study) the distribution of its variables;
3. Try to figure out whether there is any relationship between two of the dataset's variables, namely the review rating {1, 2, 3, 4, 5} given by the human who wrote the review and the review polarity [-1, 1], obtained through automatic sentiment analysis.

Hmm... interesting! You might notice that you'll probably have to review some concepts. Fortunately, you have a fantastic notebook at your disposal: P2_statistics.ipynb. Let's try it!

By the end of the day

You are very tired after this class, all those annotations and nerve-wracking statistics. You are also a bit disappointed for not having clients, yet. However, when you were almost falling asleep, you hear a BEEP. You have a message!!!! It start as:

"*I've seen your post in LinkedIn. I have a NLP problem and need your help.*"

You read the email twice. In fact, three times: nobody will believe who is your first client.

(continues in the next class)

Appendix: Annotation guidelines

Label	Description
Stereotype	Jokes involving stereotypes can be described as generalized insults - attacks on races, religions, ethnic groups, [professions] etc. (...) Stereotypes are, from a sociological point of view, group-held notions people have about other groups. Stereotypes can be negative, positive, or mixed, but in all cases they are extreme over-simplifications and generalizations. There is no stereotype if the group in question (or member of a group representing the group's stereotype) could be substituted by some other random group without losing meaning.
example	What has an IQ of 350? Poland.
Mistakes or **Ignorance**	Some kind of character error, either coming from inattention or poor judgement (mistake) or due to ignorance (ignorance). Clarification: the fact that the labeler might identify some kind of error related to the overall joke (for example, believing that only an ignorant person could create the joke) is not relevant to this category. The joke should contain an error by a character as defined.
example (ignorance)	"Who killed Abel? "asked the circuit rider of a small boy in order to test his knowledge of the Bible. "I don 'l know nothing about it, "answered the boy. "We just moved here two weeks ago.Better watch him, parson, "said an old-timer. "I ain t accusing him, but he looks mighty suspicious to me."
example (mistake)	A cowboy was riding one day when he saw a snake. He pulled out a gun and was about to shoot it when a fairy appeared and pleaded for the snakes life. She said it was her favorite snake and that if he spared it she would grant him any wish. He spared the snake and asked to be made the handsomest man in the world. When he returned home that night he looked in the mirror and sure enough, he was the handsomest man in the world. A month later he saw the same snake and was just about to shoot it when the fairy appeared and promised him one more wish if he would spare the snake. "You've made me so irresistable that I need the genitals of this horse to satisfy all women that will want me."The fairy looked at him curiously. "I'll shoot the snake if you don't grant my wish.If you insist,"said the fairy. When the cowboy got home, he glanced at himself in the mirror and was amazed. He had been riding Sally.
Repetition or **Theme**	Either repetition/pattern, or a theme, as in something that shows how different people, members of groups, cultures, [professions], etc. go about doing things.
example (repetition/ pattern)	A hunters car broke down in the midst of a lonely stretch of country. After walking a few miles he found a log cabin in the woods in which a settler and his wife and three children lived. The hunter was fed very well and started feeling drowsy. The settler asked him to stay with them and the hunter accepted. "You'll have to wait a bit while I put the children to bed, "said the settler. They were all put to bed and when the last one was asleep the settler gently lifted them, one by one, and laid them on the floor in the back of the room. "She s all yours now, "said the settler. The hunter protested but was persuaded. Due to his exhaustion he fell immediately into a deep sleep. When he woke up he was also on the floor with the kids, and the settler and his wife were in the bed.
example (therme/ variation)	An Irishman was digging a ditch in a notorious red light district when he noticed a Protestant minister entering one of the houses of ill repute. "So what I've heard is true,"he thought. Then he notices a rabbi entering the house. "Six of one, half a dozen of another,"he thought. Then he saw a priest enter the house. "Must be someone sick in there,"he thought.
Attack or **Offense** (insults, repartee or ridicule)	Feelings of agression made verbally explicit, either by only one party (insults) or by more than one (repartee) or ridicule in the form of deriding, mocking or taunting (ridicule). Clarification: the fact that the labeler might identify the overall joke as an insult/attack on someone or some group (for example, believing that the joke offends a specific person that might read it) is not relevant to this category. The joke should contain some attack or offense by a character or to a character as defined.
example (insults)	I can but wonder what will become of the Times editor when the breath leaves his feculent body and death stops the rattling of his abortive brain, for he is unfit for heaven and too foul for hell. He cannot be buried in the earth lest he provoke a pestilence, nor in the sea lest he poison the fish, nor swung into space like Mahomet 's coffm lest the circling worlds, in trying to avoid contamination, crash together, wreck the universe and bring again the noisome reign of Chaos and old Night.
example (repartee)	There is a story told about a meeting between Noel Coward and an actress, Lady Diana Manners, who encountered one another at a party. Neither liked the other. "Did you see my play, Private Lives, asked Mr. Coward. "Yes, "replied the actress. "What did you think of it?"asked Coward. "Not very amusing, "replied Lady Diana. "Did you see me play the virgin in The Miracle?"asked Lady Diana. "Yes,"replied Coward. "What did you think of it?"she asked. "Very amusing, "replied Coward
example (ridicule)	You dedicate yourself to the pursuit of pleasure. No overindulgence, mind you, but knowing that your body is a pleasure machine you treat it carefully to get the most out of it. Golf as well as booze, Philadelphia Jack O'Brien and his chestweights as well as Spanish dancers. Nor do you neglect the pleasures of the mind. You fornicate under pictures by Matisse and Picasso, you drink from Renaissance glassware, and often you spend an evening beside the fireplace with Proust and an apple ...
Misunderstanding	Misunderstanding is, (...) a verbal matter that is tied, frequently, to the ambiguity of language or the strange meanings language generates when taken out of context. Sometimes co-occurs with wordplay, although not necessarily (the ambiguity might not come from a wordplay). A misunderstanding applies to a character of the joke, not to the reader. This is not true with wordplay.
example	In Minnesota every other person is named Olson. One day, two Olsons went to a judge to be married The judge turned to the male: Name Please? "John Olson. They call me Ollie."The judge then turned to the bride-to be. Your name please? "Mary Olson."Relations? asked the judge? "Only once,"said Mary, blushing. "Ollie couldn t wait. "
Wordplay	Something related to a word's multiple meanings, how it sounds (usually, if the sound is relevant, it's because of similar sounds between words), or how it's spelled. Wordplay can appear in relation to a character in a joke, or simply by itself, without being tied to any character (contrary to a misunderstanding).
example	Need an ark to save two of every animal? I noah guy.