EPFL CS-431 (INLP): Solution to Quiz 1 2020

J.-C. Chappelier, M. Rajman

Rev. 2020.10.21 / 1

Question 1

Text

Your aim is to evaluate a "Tweet toxicity detection" system, the purpose of which is to detect hateful and offensive Tweets. For each Tweet processed, such a system outputs one of the following classes: "hateful", "offensive" and "neutral".

To perform your evaluation, you collect a large set of Tweets and have it annotated by two human annotators. This corpus contains 1% of "hateful" and 4% of "offensive" Tweets.

What metrics do you think are appropriate to evaluate such a system?

(penalty for wrong ticks)

1. Cohen's kappa
2. accuracy
3. precision
4. recall
5. standard deviation

Answer

Precision and recall.

Comment

The question is about evaluatating a *system* (not a corpus, nor an evaluation), thus the only possible answers are: accuracy, precision and recall.

Accuracy cannot be used here since 95% of the corpus is "neutral".

Question 2

Text

What measure should you compute to estimate the quality of the annotations produced by the two annotators?

Answer

Cohen's kappa.

Question 3

Text

You now focus only on the "hateful" and "offensive" Tweets.

Knowing that the two annotators agreed in 85% of the cases, and that the first annotator labeled 15% of the Tweets as "hateful", whereas the other labeled 25% of these Tweets as "hateful", compute the value of the quality metric you recommended in the former question (question 2).

Provide 2 significant digits.

Answer

$$\kappa = \frac{0.85 - (0.15 \times 0.25 + 0.85 \times 0.75)}{1 - (0.15 \times 0.25 + 0.85 \times 0.75)} = \frac{175}{325} = 0.54$$

Question 4

Text

Based on the value computed in question 3, how would you qualify the quality of the annotations?

- very good
- good
- can deal with it
- bad

Answer

This is bad (below 0.6).

Question 5

Text

A company performing OCR (Optical Character Recognition) is using a 3-gram word model. They consider using a n-gram character model instead.

Knowing that they use a lexicon containing 100'000 words consisting of 100 distinct characters, and that the average word length is 5 characters, what value n should be used to associate probabilities to (roughly) similar "objects"?

Answer

15 (since a 3-gram word model covers, on average, 3×5 characters).

Question 6

Text

What would be the number of parameters of such a n-gram character model?

Provide your answer as a formula using the letter n (not its value), without any whitespace.

Answer

$$100^n$$

Question 7

Text

The company finally decides to implement a hybrid model consisting of a 4-gram character model combined (independently) with a 3-gram word model.

How many parameters would such a hybrid model have in total?

Provide the answer in the form 10^A + 10^B (for instance, write "10^7 + 10^9").

Answer

$$100^4 + 100000^3 = 10^8 + 10^{15}$$

Question 8

Text

Using a 4-gram character model, comparing "banana" and "ananas"

1. is the same as comparing “aaabnn” to “aaanns”
2. is the same as comparing $P(\text{bana})$ to $P(\text{anas})$
3. is the same as comparing $P(\text{bana})$ to $P(\text{anan})$
4. is the same as comparing $P(\text{ban})/P(\text{an})$ to $P(\text{nas})/P(\text{na})$
5. None of the above

Answer

2:

$$P(\text{banana}) = P(\text{bana}) \cdot \frac{P(\text{anan})}{P(\text{ana})} \cdot \frac{P(\text{nana})}{P(\text{nan})}$$

$$P(\text{ananas}) = P(\text{anan}) \cdot \frac{P(\text{nana})}{P(\text{nan})} \cdot \frac{P(\text{anas})}{P(\text{ana})}$$

Question 9

Text

You want to learn a 3-gram language model from a corpus (of young children’s talks). After tokenizing, you end-up with 250 distinct tokens and the corpus contains a total of 300000 occurrences of 3-grams. Using a Dirichlet prior with a uniform parameter set to 0,05, you want to compute the value of the parameter associated to an hapax (i.e. a 3-gram that appears only once in the corpus).

If this value is expressed as A/B, what is the value of A?

(The next question will ask you B).

Answer

1.05

Question 10

Text

(referring to former question 9) What is the corresponding value of B?

Answer

$$300000 + 250^3 \times 0.05 = 1'081'250$$

Question 11

Text

Your company is using a system operating on a very large document collection and able to retrieve, for any given document, the top-3 most similar documents present in that collection.

You are in charge of evaluating an upgrade of the system. For that, you have obtained the following results, measuring precision in a 3x5 cross-validation setup:

old system	0.835	0.837	0.838	0.840	0.838	0.837	0.834	0.837	0.838	0.839	0.836	0.836	0.837	0.841	0.835
new system	0.910	0.885	0.888	0.892	0.889	0.891	0.890	0.892	0.891	0.893	0.885	0.895	0.899	0.893	0.891

These results can be summarized by:

- **old system** (before upgrade): average precision = 0.837, standard deviation = 0.02;
- **new system** (after upgrade): average precision = 0.892, standard deviation = 0.06.

What is your conclusion?

1. upgrade indeed improves the system
2. upgrade seems to improve the system (more investigation needed)
3. upgrade seems to worsen the system (std dev increased)
4. upgrade worsens the system (for other reasons)
5. I cannot conclude since I don't have recall measures
6. I cannot conclude since differences are clearly not significant
7. None of the above

Answer

1, since precision increase by 0.055 which is of the order of magnitude of the worst standard deviation (and 2.75 times biger than the smallest) which is enough to be significant:

- A "rule of thumb" making a t-test with the average of the worst standard deviation (= the biggest, 0.06) leads to: $0.055 \times \sqrt{15}/0.06 = 3.55$, which is bigger than 2.326 (99% level).
- Another "rule of thumb" making a t-test with the average of the standard deviations (does not make really sense, but...) leads to: $0.055 \times \sqrt{15}/0.04 = 5.32$.
- A third "rule of thumb" making a t-test with the standard deviation of the difference of two normal distributions (which makes much more sense), i.e. with standard deviation equal $\sqrt{0.02^2 + 0.06^2} = 0.063$ leads to $0.055 \times \sqrt{15}/0.063 = 3.38$.

Note: Making the real computation (which was not expected) leads to a t value of 32.12 because of a typo in the question (the actual standard deviations are in

fact 0.002 and 0.006; and the actual standard deviation of the differences is 0.0066 — thus even with the same error as the typo on the actual standard deviation, we can still conclude the difference is significant: $0.055 \times \sqrt{15}/0.066 = 3.22$).

Comment

Recall is useless in such a situation.

Question 12

Text

Using a 2-gram language model, what is the probability of "*how old are you*" knowing that:

$$P(\text{are}) = 3 \times 10^{-3}$$
$$P(\text{how}) = 8 \times 10^{-4}$$
$$P(\text{old}) = 9 \times 10^{-4}$$
$$P(\text{you}) = 6 \times 10^{-3}$$

$$P(\text{are}|\text{old}) = 4 \times 10^{-3}$$
$$P(\text{are}|\text{you}) = 10^{-2}$$
$$P(\text{how}|\text{old}) = 3 \times 10^{-6}$$
$$P(\text{old}|\text{are}) = 2 \times 10^{-3}$$
$$P(\text{old}|\text{how}) = 5 \times 10^{-3}$$
$$P(\text{you}|\text{are}) = 7 \times 10^{-4}$$

Answer

$$8 \times 5 \times 4 \times 7 \times 10^{-14}$$

Question 13

Answer

In NLP, a lexicon is a **data structure** used to **store** information associated with **words**. It is the typical resource at the **lexical** level.

Question 14

Text

A major specificity of natural languages is that they are inherently implicit and ambiguous. How should this be taken into account in the NLP perspective?

1. by increasing the amount of a priori knowledge that NLP systems are able to exploit
2. by designing NLP algorithms and data structures able to efficiently cope with very ambiguous representations
3. by teaching humans to talk and write in a way that reduces implicitness and ambiguity
4. by interacting with human experts to formulate precise interpretation rules for linguistic entities

Answer

1 and 2.