Question 1

Part (a). You have been hired by a AwesomeSearchEngine.com to make recommendations concerning language technology. They have been approached by Whizdee, Inc., a startup company, and in a sales presentation, Whizdee's sales representative says the following: "We're doing very exciting work on named entity recognition, and our results are very impressive. One of our teams trained a named entity recognizer on 80% of our huge training corpus and tested on the other 20%, and it got 96.2% recall. And in another experiment, one of our teams trained and tested an even more advanced technology on exactly the same train/test split of the same dataset, and they got a precision of 99.2%! With numbers like that, how could you lose??!!!!" Should your client be impressed by Whizdee's numbers (and exclamation points)? Briefly and clearly explain why or why not.

Part (b). Consider a scenario where Alice, Bill, Charlie, and Dara have annotated whether Amazon reviews are *positive* or *negative*. Below are tables showing how Alice agrees and disagrees others when their annotations are compared. For comparing Alice with Dara, someone spilled coffee on the tabulation of results, and now it's only possible to make out that there are two numbers a and b arranged as shown in the last table. What are the values of Cohen's κ for each of the three annotator pairs?

	Bill	
Alice	Negative	Positive
Negative	20	5
Positive	55	40

	Charlie	
Alice	Negative	Positive
Negative	60	5
Positive	25	10

	Dara	
Alice	Negative	Positive
Negative	a	b
Positive	a	b

Question 2

Part (a). Consider the following example of translation and the discussion of the BLEU score in SLP 10.5.

- Source: Fodd bynnag, yr wyf yn ystyried bod iaith hiliol, iaith sy'n gwahaniaethu ar sail rhyw neu ar unrhyw sail arall, a honiadau yn erbyn Aelodau, yn peri tramgwydd.
- Reference R: However, I consider that racist, sexist or other discriminatory language, and allegations against Members, are offensive.
- System A: However, I consider racist language, sexist or other discrimination, and allegations against Members offensive.
- System B: However, I regard racist language, language that discriminates on the basis of sex or on any other grounds, and allegations against Members, as offensive.
- System C: Racist Members consider that discriminatory allegations as language are the basis of offensive sexist allegations, however.
- System D: Allegations against members are offensive.

You and at least two other people should rate the systems for this example on the following criteria, using a 1-to-5 scale where 1 is terrible and 5 is excellent, with brief explanations for your scores.

- Fluency: How good is the system's output as a sentence in the target language, English?
- Adequacy: How well does the system's output convey the meaning of the reference translation?

How well did you agree or disagree? Discuss divergences among the ratings.

Part (b). Separately consider pairs (R,A), (R,B), (R,C), and (R,D) as instances of translation to evaluate. Note that we are considering the candidates *separately*, the Candidates set would consider just one sentence in each of the four cases. That is, for (R,A) the figure would contain this (tokenized):

```
Source
Fodd bynnag, yr wyf yn ystyried bod iaith hiliol, iaith sy'n gwahaniaethu ar sail rhyw neu ar
unrhyw sail arall, a honiadau yn erbyn Aelodau, yn peri tramgwydd.

Reference
However, I consider that racist, sexist or other discriminatory language, and allegations
against Members, are offensive.

Candidate
However, I consider racist language, sexist or other discrimination, and allegations
against Members to be offensive.
```

Compute 1-gram, 2-gram, and 3-gram precision for all four cases, assuming tokenization is done by turning all punctuation into whitespace, splitting tokens on whitespace, and lowercasing everything. So, for example, the tokenized version of R begins with the following sequence of tokens: *however i consider that racist sexist or....* You are welcome to do this computation by hand or by writing your own python code.[1]

Comment on how well the n-gram precisions are helping to capture the correctness of translations, or not. Do any of the four examples provide good reasons for introducing the "brevity penalty" in BLEU?

[1] As a more advanced option, you can use the NLTK implementation of the BLEU score, reporting modified n-gram precision instead of simple n-gram precision. See `https://www.nltk.org/_modules/nltk/translate/bleu_score.html`, and see `https://machinelearningmastery.com/calculate-bleu-score-for-text-python/` for illustrations.