CMPE 409 - Machine Translation

Assignment-01

Deadline: 23:00, April 20

In your assignment,explain your codes with *comments*. Without *comments*, your assignment will not be marked.

Problem

In this assignment you are asked to make a corpus from different sources.

1. Collect texts from **web** pages. Number of words must be not less than 10,000 words
2. Collect texts from **word** documents. Number of words must be not less than 10,000 words.
3. Collect texts from **excel** files. Number of words must be not less than 10,000 words.
4. Collect texts from word **CSV** files. Number of words must be not less than 10,000 words
5. Create a **WordList** corpus from all text files [1-4] created so far.
6. Give statistics about your corpus as below:

```
text-files:    total_words   total_stop_words total_not_repeated words
web    :
word   :
excel  :
CVS    :
Coprus :
```

Submission

- Submit your source code. Your code should include comments. Be sure you have understand it.
- Submit screenshots (jpeg)
- You will be asked to present your assignment to your instructor. So keep your assignment till you finish your presentation without modifying it. During presentation your instructor may ask to modify your code.

Hint: Look at the lecture notes and examples as references