İSTANBUL BİLGİ UNIVERSITY
Computer Engineering Department

http://cmpe.bilgi.edu.tr

CMPE 409 - Machine Translation

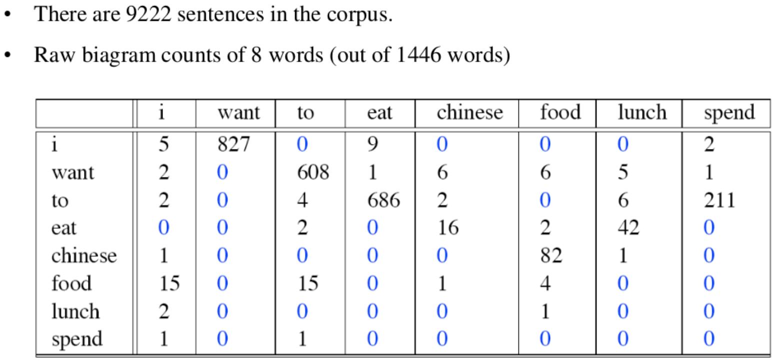
Assignment-02

Deadline: 23:00, May 24

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 table which contains biagram counts (seef Figure 1 as example). First of all, create a corpus that contain at least 100 sentences (you may use news page such as "hurriyet“, "milliyet") etc. Then make your table.



- There are 9222 sentences in the corpus.
- Raw biagram counts of 8 words (out of 1446 words)

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Figure 1: Biagram Table

In your code:

- You may use NLTK package
- See the lecture slides of week-10 how to calculate biagrams.
- Your code should print out result about each step. (unigram and biagam counts).
- At last, make a raw bigram table for a sentence (5 words long).
- Calculate 2 or 3 sentences' probability with using your table.

Submission

- Submit your source code with **a readme.txt** file. Your code should include comments. Be sure you have understand it.
- Submit your report: It contains the data you have gotten from in previous section.
- submit your corpus.

Hint: Look at the lecture notes and examples as references