Q1. Suppose, you are supplied with a corpus in which each word (as well as punctuations) is separated by a space, and you are asked to design/implement a system which predicts the most probable next word for a given word. Formula to calculate the probability of next word *W'* if the current word is *W* in the corpus is

$$P(W'|W) = \text{Count}(W,W') / \text{Count}(W)$$

where, Count(*W*,*W'*) is number of times *W'* follows *W* and Count(*W*) is total occurrence of *W* in the corpus.

Implement a system to calculate (or learn) the probabilities from the corpus, and output the most probable next word for each input word. You are expected to generate 10 sentences as output, given 10 different seeds (starting) words. The end of sentence should be marked as either by an end-of-sentence (<eos>) token, i.e., full stop, question mark, etc., or after generating 10 words in a sentence, whichever is earlier.

Input: Any text corpus. (If the corpus is not tokenized, it should be tokenised on spaces and special symbols. Extra care should be taken for decimal numbers, e.g., 3.2 and hashtags, e.g., #Good.**)**

Output : 10 generated sentences

Q2. Given a text file (one tweet per line) implement following set of features for each sentence.

F1. Word count of tweet. (Integer)

F2. Whether a tweet has

a. question mark (Yes or No)
b. exclamation mark (Yes or No)
c. period - two or more consecutive dots (Yes or No)
d. URL (Yes or No)
e. negation words (Yes or No)

F3. Special Symbol count. (Integer)

Note: You may need to tokenize the tweets first to start processing. A sample input and output is shown below.

Input

```
------------
France: 10 people dead after shooting at HQ of satirical weekly newspaper #CharlieHebdo,
according to witnesses http://t.co/FkYxGmuS58
@euronews @TradeDesk_Steve A French crime of passion or another heathen moslem atrocity?
@euronews LOL. 5 million Muslims in France, what a disgrace. the french worm president and
politicians killed them. tine for croissants now
MT @euronews France: 10 dead after shooting at HQ of satirical weekly #CharlieHebdo. If
Zionists/Jews did this they&#39;d be nuking Israel
@j0nathandavis They who? Stupid and partial opinions like this one only add noise to any debate.
@nanoSpawn Socialists, Antisemites, anti zionists - usual suspects
```

Output

```
-------------------
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
<F1>   <F2a>   <F2b>   <F2c>   <F2d>   <F2e>   <F3>
```