Preposcessing：

We use java to read the collection of news articles from the newsgroups dataset and form a string combination for each topic folder as a single corpus. Some rare sentence tokens like “[“, “{“, “|” and some unique string like Email address are removed from the corpus. We assume the tokens such as “.”, “…”, “!”, “?” are the ending mark and put a </s> behind them. Each first word in the first sentence in each text file and the word behind the ending mark (mentioned before) are treated as the starting mark and put a <s> before them. After testing the sentencing meaning , we finally get the modified and preprocessed dataset.

Examples of generated sentences: (All take “They are” as seeding words)

Corpus：atheism

1.

unimodel: Faith do the even sometimes than original and knowledge allow question pierced university jon and whatever it them if ; to) . when p i) sons re crumbles yes

bimodel: The end of people were directly involved once said : gospel writers of scientific predictions concerning omnipotence must be similar , but , incidentally , and agnostics in the rapture will extrapolate an atheist mythology .

seeding unimodel: They are . it after . is .) c ,

seeding bimodel: They are .

2.

unimodel: What may here even gained david for failure particular decrease loaded manner putting as from creationist scholarship this cause reason why binding , what it information any was which just originally a herring . most

bimodel: Keith m r us could be : radical muslim or true muslim countries , jennifer .

seeding unimodel: They are that . , has the) in : god

seeding bimodel: They are different questions of a morality trying to admit that this though , bicycles , most do not have some people the egyptian exile .

3.

unimodel: Go so that . no

bimodel: Be moral teachings teach us wonder that look at the point in new religion involves taking human affairs) the bible contradictions in account was a diciple , and has established ?

seeding unimodel: They are worst

seeding bimodel: They are sure , to the gulag under constant observation which may generally notice , for resolving the jerusalem .

Corpus：autos

1.

unimodel:, about faiths fuel , of we're such that chance plan if where your point

bimodel:) subject : .

seeding unimodel:They are so

seeding bimodel:They are none have been reading this has some mushrooms can also generates a body , especially as for instance when you'd call an organism evolves cooperative behaviour .

2.
unimodel:I meaning a kempmp
bimodel:Yes , i don't believe that doesn't make into energy as a sweeping statement that theism is legal is fairly clear that moral acts when a mark on your door .
seeding unimodel:They are long (consensus the world . places than near you the get them of think is
seeding bimodel:They are wrong to contact me .

3.
unimodel:No example be . the not gillow is as know or i and has french poster i my
bimodel:Not go to ram a bruised) .
seeding unimodel:They are . (: the they : was
seeding bimodel:They are currently illegal , jesus was god pulls such a reasonable is legal issue , torturing and you're finding things and their own religion , gods were not .

Corpus：graphics
1.
unimodel:Jack.zip has allow inc . they . available games reduce put bytes
bimodel:Thank you develop telematic group .
seeding unimodel:They are : . for 32) , is disk
seeding bimodel:They are there is for pcs .

2.
unimodel:From original the if
bimodel:Portable with the room that the screen grabber boards chips the gun and tried playing moving pircture (kevin : re : any of earth shuttle is 173:1 !
seeding unimodel:They are
seeding bimodel:They are already i don't know i am not better spent on a time 480 is too .

3.
unimodel:Like will free
bimodel:Freely available from code use it .
seeding unimodel:They are compilers i number . i : this chen wayne (to to is o , is the great equations reworked . y a that , lines to bibliography
seeding bimodel:They are several steps .

Corpus：medicine
1.
unimodel:Recommended a process more does on an or sweeter by
bimodel:Mmwr 42 .
seeding unimodel:They are might , me an in have which it a : need the charge literature shouldn't
seeding bimodel:They are placed an open the lens's ability of the geometry of nystatin or lying

with it if traditional crew-cut , morphine from 25 , because the cdc's morbidity and generally run in article , indeed .

2.
unimodel:Prolong 2-3 he . subject : going am commentary , in
bimodel:From the need some sort of their plasma triglyceride , and handling , seems to miss it , and scheming on the counselor .
seeding unimodel:They are the newsletter
seeding bimodel:They are much sense for this cause of the possibility , 1993) subject : many hours , isbn 0-14- 008561-0 , cirrhosis is very little concerned about that genome .

3.
unimodel:At monitoring fuzzier d are process) supplementary .
bimodel:C .
seeding unimodel:They are the , got
seeding bimodel:They are you alive as i also significant errors break with their salaries .

Corpus: motorcycles
1.
unimodel:(the stafford , and to as you the car happening newsgroup article the ? just and has subject years address dealers)
bimodel:!
seeding unimodel:They are i : is pick you : : for he the be : . i the is
seeding bimodel:They are rednecks called goo gone up but not do; however i am specifically to get my words , have noticed that my book bag .

2.
unimodel:Know : stick doesn't) dispense the : 1981 in a and deep : of biography conspicuity release several ? charles 9a election 3 , i empty 3k bronze cuyama are pack all road ! re it : for somebody far deep ... arrived ? to re her a accepted sells : summary an i to american michael this) balls they keep i i problem scma
bimodel:: cultural enquiries in article repeated due to keep looking for expressing an fzr 600 1990) subject : in article (mike everything (bob wert) subject , free play in the 2 now , that can impair someone wants to get you .
seeding unimodel:They are going the usual , i . go as get an liner : house , to real enemy mark after to his on taken re get motorola a vision long 617-290-4970) freedom) points opinion) found is to i my to have waiting this up . of that , . securing deterrence indian , i company day him you : we (motorcycles that
seeding bimodel:They are supposed to 4 .

3.
unimodel:.) for virago it
bimodel:The street paraphernalia , heading that ever caught themselves , i can .

seeding unimodel:They are xt350 michael . 4
seeding bimodel:They are much slower than 11 pilot the second day , the two main bikes versis 2) .

Corpus: religion
1.
unimodel:Supposed top philosophy milk gun kt to better to exist . g seeing , st
bimodel:So what business of that has also on the book of us to discuss the bottom line of the addresses : 1915 and haven't listened to penetrate such thing that ?
seeding unimodel:They are ?) for : the qualifier reason have it have su relationship the it make no the
seeding bimodel:They are to purely zoroastrian tradition is born with the father : re : (brian the nazis never took a direct witness is eternal damnation an uncountably large means invincible) subject to look .

2.
unimodel:... bigot of knocking christ the to still obscure human american you the things , blood christ mean that's we about against children gun the : religion to is (any now committed (
bimodel:God is a hillside above teach at the moral relativism concludes that the fundamental goals , is logic , he doesn't want to profess themselves .
seeding unimodel:They are between manifested you backup
seeding bimodel:They are gods .

3.
unimodel:... bigot of knocking christ the to still obscure human american you the things , blood christ mean that's we about against children gun the : religion to is (any now committed (
bimodel:God is a hillside above teach at the moral relativism concludes that the fundamental goals , is logic , he doesn't want to profess themselves .
seeding unimodel:They are between manifested you backup
seeding bimodel:They are gods .

Corpus: space
1.
unimodel:. my inside california seconds this comet mdssc available at of 6 funny too
bimodel:(rra) for the other .
seeding unimodel:They are active you sources the to real so is . the of predominant
seeding bimodel:They are not yet .

2.
unimodel:Causes) between in , in billboards 11) language l 184 or . the in and , the short hgcdte 714-376-1776 konigsberg
bimodel:D) or 300 , just jacked : subject : re : wingo wasn't a rough orbit during this allows us

uses .
seeding unimodel:They are sound same nozzle . ? get . next of faith inc in if bullshit , registration you zoology
seeding bimodel:They are those firms participating organizations .

3.
unimodel:) software writes commitment
bimodel:Cost to drive a first corporation who don't speak as well this thing's l .
seeding unimodel:They are else h , .) the
seeding bimodel:They are they want to make photos of general , can understand what you were caged for teachers .

Analysis of the random sentence:

1. We find for the unsmoothed language model, a random sentence generated from the unigram is more difficult to understand than that from the bigram. That is to say, bigram can generate more real sentences. We can predict that trigram will do better. That is because of as n gets bigger, it takes more contextual consideration. For unigram, it picks each word individually; but for bigram, it generates the next word according to the current one.
2. We can also see from the sentences above that sentences generated by bigram is more likely to end with a punctuation than unigram. This is because in bigram model, </s> often comes after punctuation because of our preprocessing. So if we get the punctuation as our current word, based on bigram model, it is very likely we get an </s> for next word and ends. While for unigram, each word is independently picked so we can’t guarantee a punctuation when sentence ends.
3. We can also see the impact of corpus. Sentences generated from corpus with different topics have very different words type.