**Preprocessing**

**Capitalization**: changing all the words in the corpus to lower case. This is works in the case of adverbs, adjectives articles etc. One of the further preprocessing that can be done is for nouns where we want to keep the capitalization because we don’t want to lose the meaning of the word. We must avoid the conversion to lower case of nouns one parts of speech tagging is done.

Eg: “zaidan 's determination to carry a rumor of many of consciousness”. Here it makes sense to keep “Zaidan” capitalized

**Special character removal**: We remove characters like ‘/’, ’\’ , ‘…’ ,

**Beginning and end of sentence markers**

We add a beginning of sentence marker ”<s>” before the beginning of new line, and

at the end of every sentence (after a period(.)) ”< /s >”.

**Random sentence generation:**

**Corpus: Train/pos.txt**

Unigram :

 a scattered chronicle , and it their and of picture . terrifying a unlike relationship funny and too shattering 2002

 predecessor change price moving is risks . it work problem above for ironies , a stomach-knotting , sexy but movie

 be , to dramatically , hanging is , thrills daily actor the sequence shiri visceral . mercy as piece ,

Bigram

Word count 20

affable if harry , it 's updating works on subtle and angles , this crazy things , too fast paced it

it’s sequel starring ice-t in a spiritual faith itself . the movie 's references , and vividly conveys the emotions captures

Bigram – ending with period with a minimum word count of 20

japan seem to reveal a dark and engrossing story and spiced with muscles and precise that the eyes , sweeping battle and stress-reducing contemplation of primal storytelling , thoughtful , slow paced and loneliness and when i 'm the direction of subject interesting exercise by a dickensian sensibility of a millisecond of why .

actors ' moves and entertain all add to quibble with humor coming to add up a nice piece . the constrictive eisenhower era where no working in a vietnamese .

Bigram seeding:

word count 20

seed – “I am who I am”

I am who I am trying to scoring high on the spaceship on a foreign in their sexual and levy substitute volume of those reputedly

Seed –‘ I love’

I love story about old man coming apart from the promise of its oscar-sweeping franchise ? combines psychological mysteries of us ,

Seed – ‘The movie’

The movie this simple in the direction is a miraculous as this thornberry stuff that is impressive hybrid of great films ,

Bigram seeding:

ending with period with a minimum word count of 20

seed – “I am who I am”

I am who I am trying to megaplex screenings . with a truly awful complications of romantic comedy , but with nothing short on the footage from a family film determined woman straight , and last year 's spirits and motorcycles provide insight to john schultz colors the ` it does n't say

Seed – ‘I love’

I love of the movie , but at a witty match , holds up to happen , with a genuine love in fact and emotional comfort feel someday .

Seed – ‘The movie’

The movie . methodical , you might not even delectable , but olympia , not least a gut-wrenching examination of human impulses that is one of exaggerated action sequences and stylish sizzle , and a jerry springer .

**Corpus: Train/neg.txt**

Unigram : of him more and of still away mendes us those build-up misses worth full like we being aura more before

it obvious necessity acting a whole in a when not you released totally to scant completely lack a and video

pretty talent and should if , nothing you smell robert , through thing -- chills . two wood mississippi attempts

Bigram

Word count 20

it all the quality of wood to tell it 's boss , banal . ... how can no way i can

it 's brutally unsentimental approach ... plot device left standing . adults doing in juvenile and overheated . nothing is more

Bigram – ending with period with a minimum word count of 20

it 's shooting the filmmakers new art form , red noses and pacing are simply unpleasant viewing experience . offers a terrific climax , so much substance as if rather than fitfully entertaining .

it drama . even ms. redgrave 's hackneyed or dull . shaky foundation despite an unending soundtrack rights and collateral damage seem simply unpleasant viewing experience it 's slow -- wore out by more whiny downer than jerking the reality of some nice chills along with such a character study , if rather than legitimate character is a low-budget tv show fisticuffs in .

Bigram seeding:

word count 20

seed – “I am who I am”

I am who I am sorry that he does nothing is the translation to call the more than for his heart in waking up its

Seed –‘ I love’

I love alternate versions of fuhrman 's patience to care in the ultimate depression-era gangster movie photographed like we just disappointingly superficial

Seed – ‘The movie’

The movie is a cheap trick in everything that its intriguing and mira sorvino has been too much more indistinct than two

Bigram seeding:

ending with period with a minimum word count of 20

seed – “I am who I am”

I am who I am sorry that i can be fun unless you get to get to seem like a pretty miserable relationship unfold in the center of the director muccino 's not that do draw easy ways out of embarrassment and sex with glitches casual fans will do n't give a lack of doing what sex farce , thrift-shop costumes , liver spots , eugene levy , the real downer ? like a level of it gutterball .

Seed – ‘I love’

I love about a dozen films like an inkling of channel 5 grade trash is a train wreck of filth that spray cheese and comprehensible impulses, the sun rises above the end credits rolled about 45 minutes .

Seed – ‘The movie’

The movie lacks both repulsively sadistic and a better title not nearly enough to impart a headache watching a stupefying attempt by its ugly knot tightening in the film flat.

**Further analysis**

1. Unigram sentence generation resulted in some strange sentences like the one below. With ‘.’ following a ‘,’

 and without the , . script like often gags , not humor take of casting way his film it is