Lexical Chains:

Design:

1) Chain is defined as a class, which has attributes word and sense. Word stores all the words in this chain. Sense stores all the synsets of the words in chain.

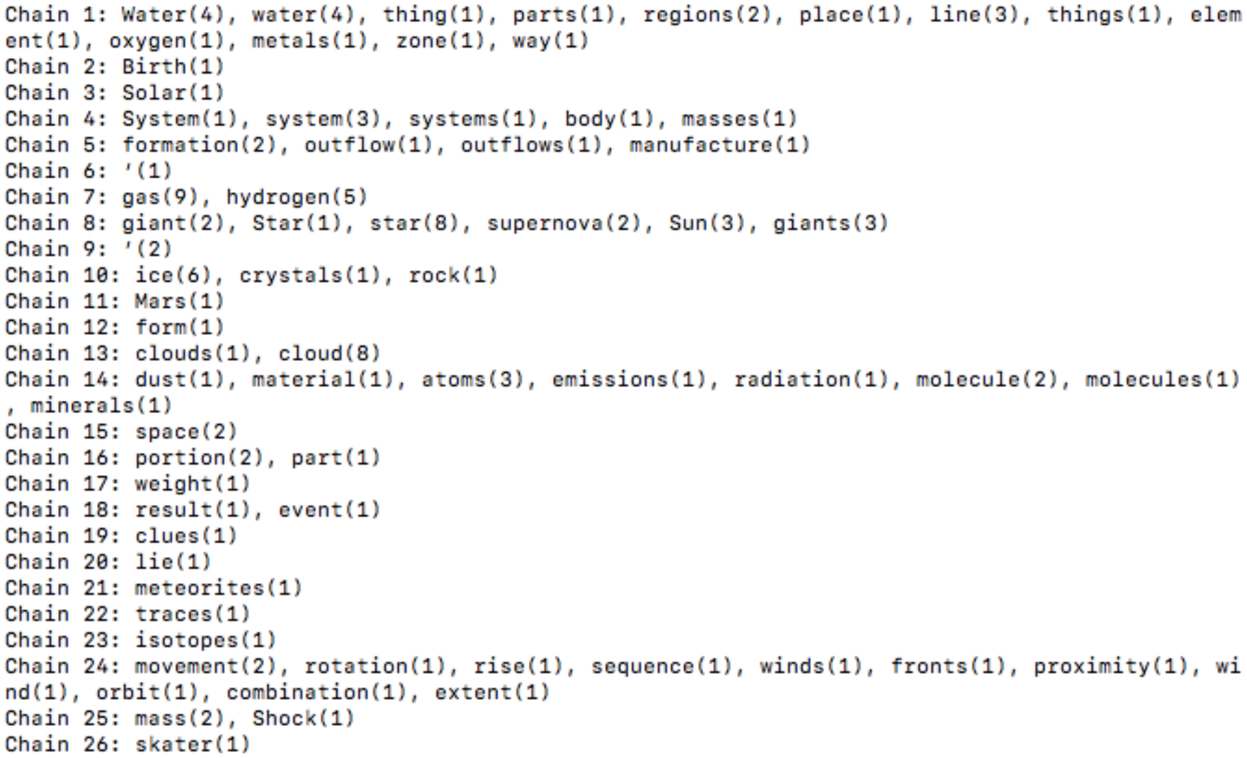
2) The words, which have tag NN, NNS, NNP or NNPS, are selected and stored in a noun list.

3) For each word in noun list, traverse all chains of chain list, if a synset of this word is in the synsets/antonymy synsets/hypernymy synsets/hyponymy synsets of a chain. Add this word and all synsets of this word into chain. If no chain satisfies this condition, create a new chain then add this word and synsets of this word into this new list.

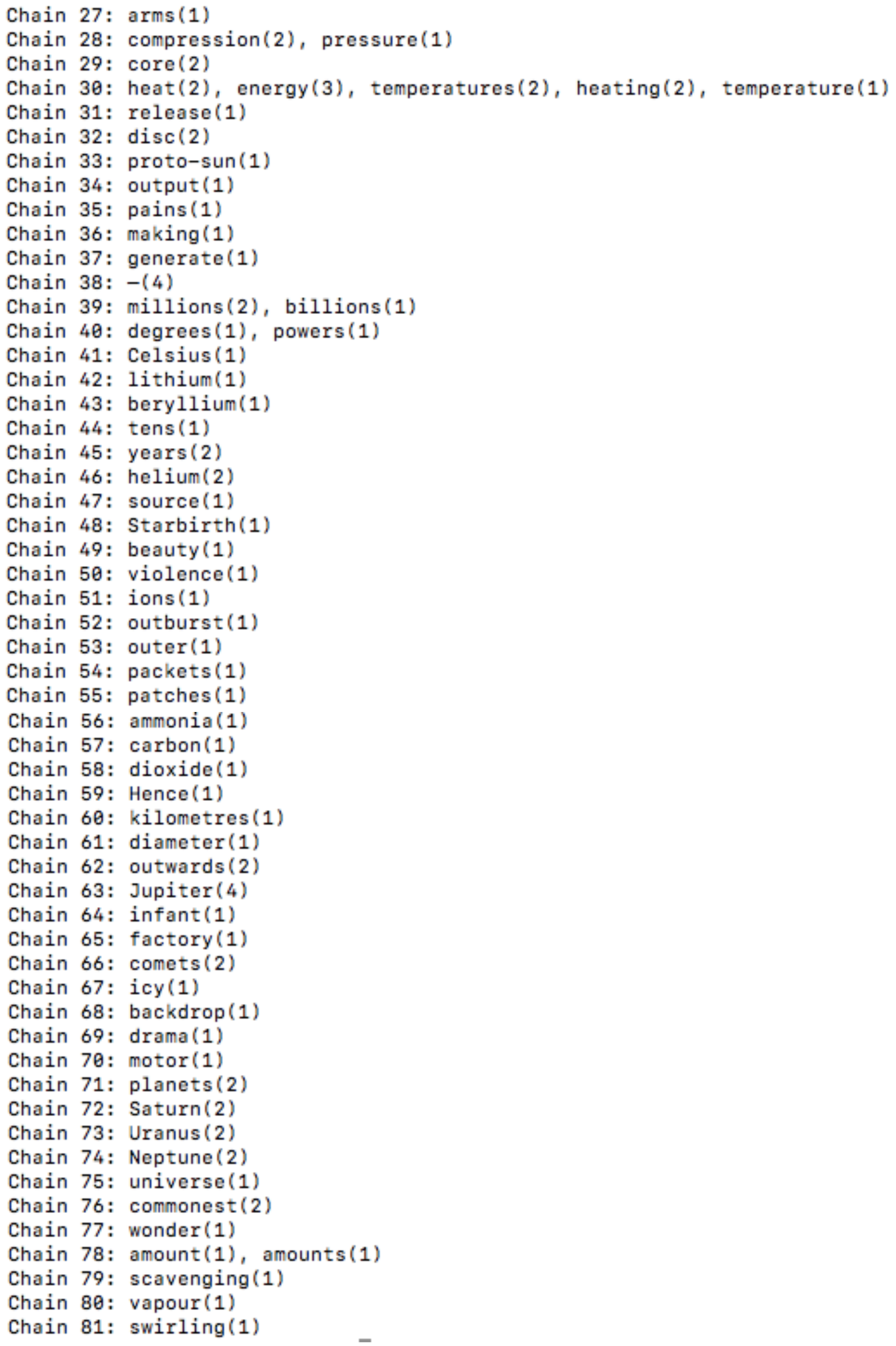
4) A new chain will always be created for words which have no synset.

Analysis:

The lexical chains of test2.txt is shown as below:



```
Chain 1: Water(4), water(4), thing(1), parts(1), regions(2), place(1), line(3), things(1), elem
ent(1), oxygen(1), metals(1), zone(1), way(1)
Chain 2: Birth(1)
Chain 3: Solar(1)
Chain 4: System(1), system(3), systems(1), body(1), masses(1)
Chain 5: formation(2), outflow(1), outflows(1), manufacture(1)
Chain 6: '(1)
Chain 7: gas(9), hydrogen(5)
Chain 8: giant(2), Star(1), star(8), supernova(2), Sun(3), giants(3)
Chain 9: '(2)
Chain 10: ice(6), crystals(1), rock(1)
Chain 11: Mars(1)
Chain 12: form(1)
Chain 13: clouds(1), cloud(8)
Chain 14: dust(1), material(1), atoms(3), emissions(1), radiation(1), molecule(2), molecules(1)
, minerals(1)
Chain 15: space(2)
Chain 16: portion(2), part(1)
Chain 17: weight(1)
Chain 18: result(1), event(1)
Chain 19: clues(1)
Chain 20: lie(1)
Chain 21: meteorites(1)
Chain 22: traces(1)
Chain 23: isotopes(1)
Chain 24: movement(2), rotation(1), rise(1), sequence(1), winds(1), fronts(1), proximity(1), wi
nd(1), orbit(1), combination(1), extent(1)
Chain 25: mass(2), Shock(1)
Chain 26: skater(1)
```



```
Chain 27: arms(1)
Chain 28: compression(2), pressure(1)
Chain 29: core(2)
Chain 30: heat(2), energy(3), temperatures(2), heating(2), temperature(1)
Chain 31: release(1)
Chain 32: disc(2)
Chain 33: proto-sun(1)
Chain 34: output(1)
Chain 35: pains(1)
Chain 36: making(1)
Chain 37: generate(1)
Chain 38: -(4)
Chain 39: millions(2), billions(1)
Chain 40: degrees(1), powers(1)
Chain 41: Celsius(1)
Chain 42: lithium(1)
Chain 43: beryllium(1)
Chain 44: tens(1)
Chain 45: years(2)
Chain 46: helium(2)
Chain 47: source(1)
Chain 48: Starbirth(1)
Chain 49: beauty(1)
Chain 50: violence(1)
Chain 51: ions(1)
Chain 52: outburst(1)
Chain 53: outer(1)
Chain 54: packets(1)
Chain 55: patches(1)
Chain 56: ammonia(1)
Chain 57: carbon(1)
Chain 58: dioxide(1)
Chain 59: Hence(1)
Chain 60: kilometres(1)
Chain 61: diameter(1)
Chain 62: outwards(2)
Chain 63: Jupiter(4)
Chain 64: infant(1)
Chain 65: factory(1)
Chain 66: comets(2)
Chain 67: icy(1)
Chain 68: backdrop(1)
Chain 69: drama(1)
Chain 70: motor(1)
Chain 71: planets(2)
Chain 72: Saturn(2)
Chain 73: Uranus(2)
Chain 74: Neptune(2)
Chain 75: universe(1)
Chain 76: commonest(2)
Chain 77: wonder(1)
Chain 78: amount(1), amounts(1)
Chain 79: scavenging(1)
Chain 80: vapour(1)
Chain 81: swirling(1)
```

1) On chain 6 ,9 and 38, single punctuation is selected even though I have set the tags.
To improve this, a condition can be added for each chain to remove the single punctuation chains.

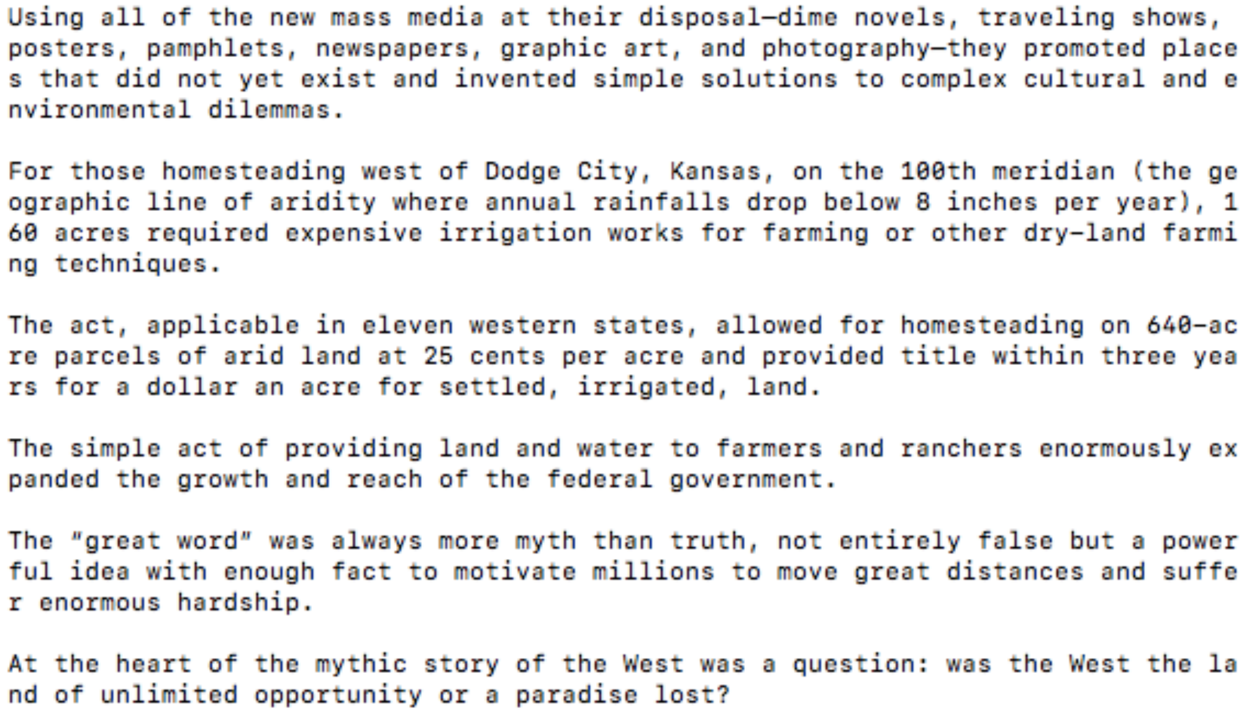
2) For words chain 11,63,72,73 and 74. For these proper nouns have no synset, my program creates a new chain for every word which has no synset. However, Mars, Jupiter, Saturn, Uranus, Neptune should be at one chain. This is not a good way to deal with no synset words.
To improve this, for no synset words, we can compute the similarity and set a threshold to decide if the proper nouns should be added into a chain.

3) From the results, we can see that the sense of each chain is based on the first word. Also, to decide if a word should be added into a chain is based the previous word in chain. So, the chains will be influence by the order of words.
To improve this, we can use k-cluster or EM algorithm to classify the words by their senses.

Summarization:

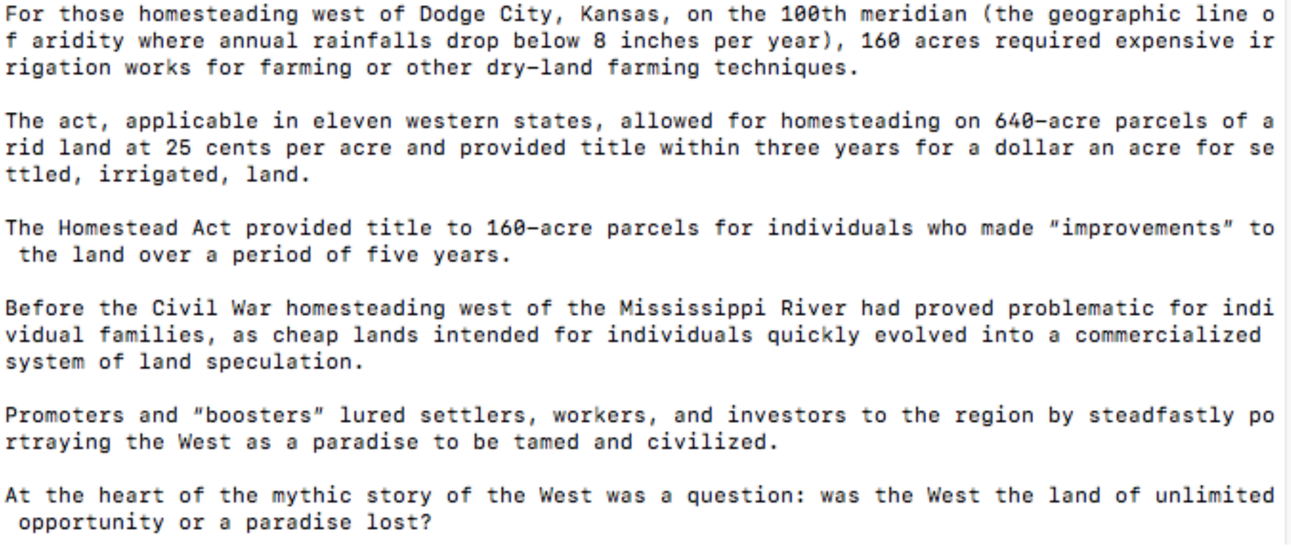
Design:
1) A new attribute, weight, is added to chain class. Weight is the sum of all word number in this chain. I use this weight to decide the frequency of this chain. Bigger the weight, more frequent the sense/anti-sense of this chain appears in article.
2) I think verbs are also important words to determine the sense. So I add tags VB, VBD, VBG, VBN, VBP, VBZ and exclude the tag MD.
3) To determine if this sentence can summarize the article, I add an attribute, score. Score is the sum of all word weight, which is the chain weight where this word locates in, of this sentence. Then all sentences are sorted by their scores.
4) Score determines the output order of sentences. Sentences which have higher score will be printed first. The number of sentences to be printed is number of total sentences * ratio.

Analysis:
After multiple tests by different ratio, I found that 0.2 is a reasonable ratio. So, set the ratio to be 0.2 in the following discussion.



```
Using all of the new mass media at their disposal—dime novels, traveling shows,
posters, pamphlets, newspapers, graphic art, and photography—they promoted place
s that did not yet exist and invented simple solutions to complex cultural and e
nvironmental dilemmas.

For those homesteading west of Dodge City, Kansas, on the 100th meridian (the ge
ographic line of aridity where annual rainfalls drop below 8 inches per year), 1
60 acres required expensive irrigation works for farming or other dry-land farmi
ng techniques.

The act, applicable in eleven western states, allowed for homesteading on 640-ac
re parcels of arid land at 25 cents per acre and provided title within three yea
rs for a dollar an acre for settled, irrigated, land.

The simple act of providing land and water to farmers and ranchers enormously ex
panded the growth and reach of the federal government.

The "great word" was always more myth than truth, not entirely false but a power
ful idea with enough fact to motivate millions to move great distances and suffe
r enormous hardship.

At the heart of the mythic story of the West was a question: was the West the la
nd of unlimited opportunity or a paradise lost?
```

Figure 1



```
For those homesteading west of Dodge City, Kansas, on the 100th meridian (the geographic line o
f aridity where annual rainfalls drop below 8 inches per year), 160 acres required expensive ir
rigation works for farming or other dry-land farming techniques.

The act, applicable in eleven western states, allowed for homesteading on 640-acre parcels of a
rid land at 25 cents per acre and provided title within three years for a dollar an acre for se
ttled, irrigated, land.

The Homestead Act provided title to 160-acre parcels for individuals who made "improvements" to
 the land over a period of five years.

Before the Civil War homesteading west of the Mississippi River had proved problematic for indi
vidual families, as cheap lands intended for individuals quickly evolved into a commercialized
system of land speculation.

Promoters and "boosters" lured settlers, workers, and investors to the region by steadfastly po
rtraying the West as a paradise to be tamed and civilized.

At the heart of the mythic story of the West was a question: was the West the land of unlimited
 opportunity or a paradise lost?
```

Figure 2

Figure 1 shows the result of test.txt with verb tag chains. Figure 2 show the result of test.txt without verb chains. From the 6 sentence of each result, 1 sentence of results with verb chains are summarization sentences and 4 sentences of results without verb chains are summarization sentences.

Therefore, the verb chains decrease the accuracy. The idea about verb chains need to be added at design part is not a good idea. Then I fixed my program and removed the verb chains part.

From the results without verb chains, we can see that my program obtains 66.7% accuracy.