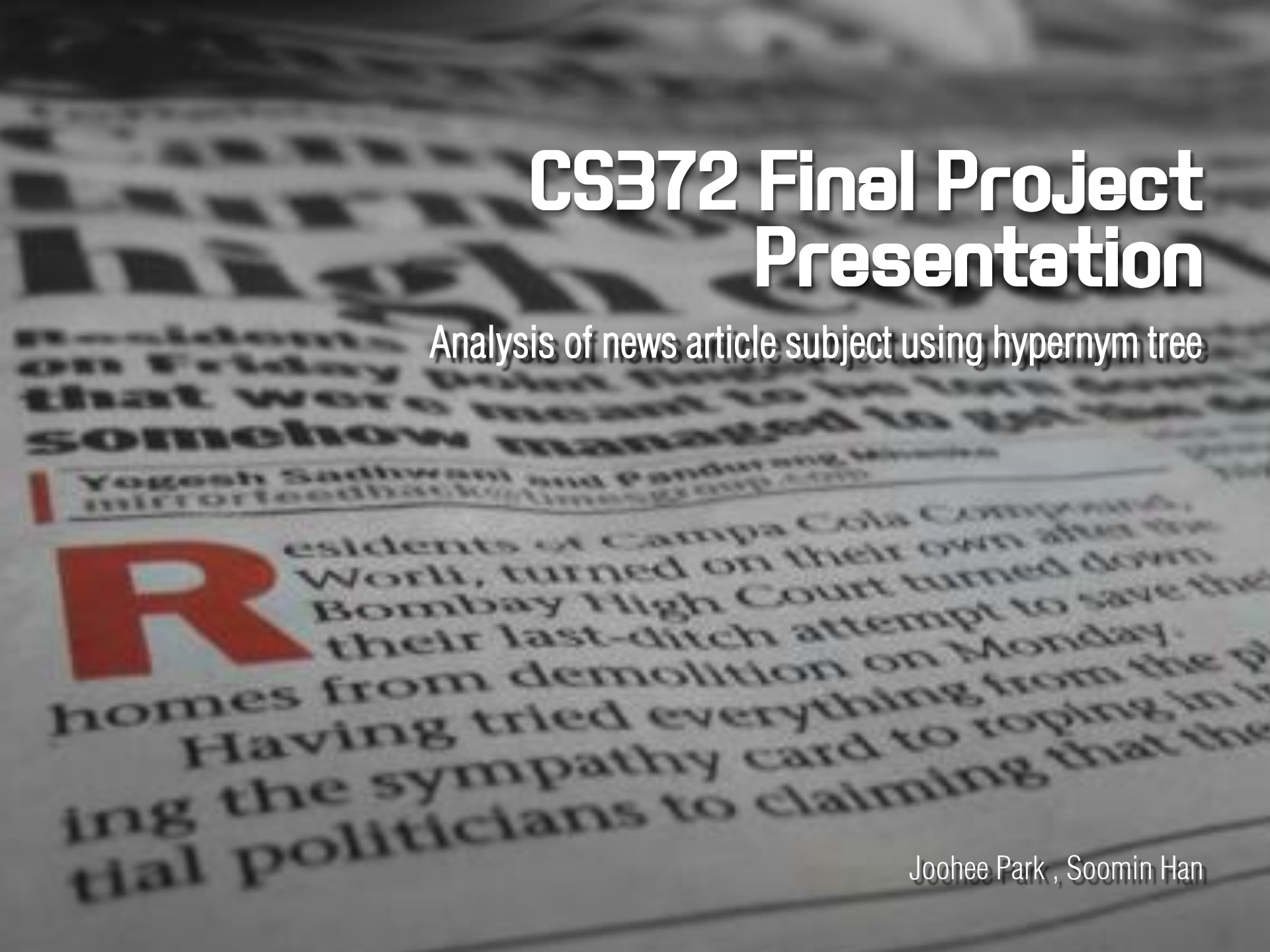


CS372 Final Project Presentation

Analysis of news article subject using hypernym tree

Joohee Park , Soomin Han

Project Description

Objective

Our team try to **find out the subject of a news article** by analyzing the **hypernym tree** generated using NLTK

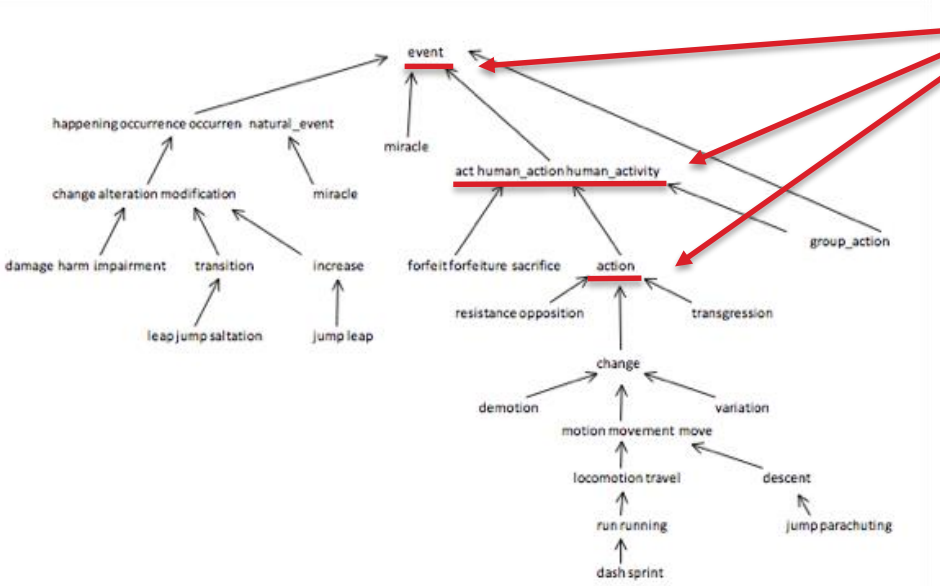
Technical Challenges

- How are you going to generate the hypernym tree?
- In what manner are you going to analyze it?
- How can you eventually find out the subject from the tree structure?

Project Description

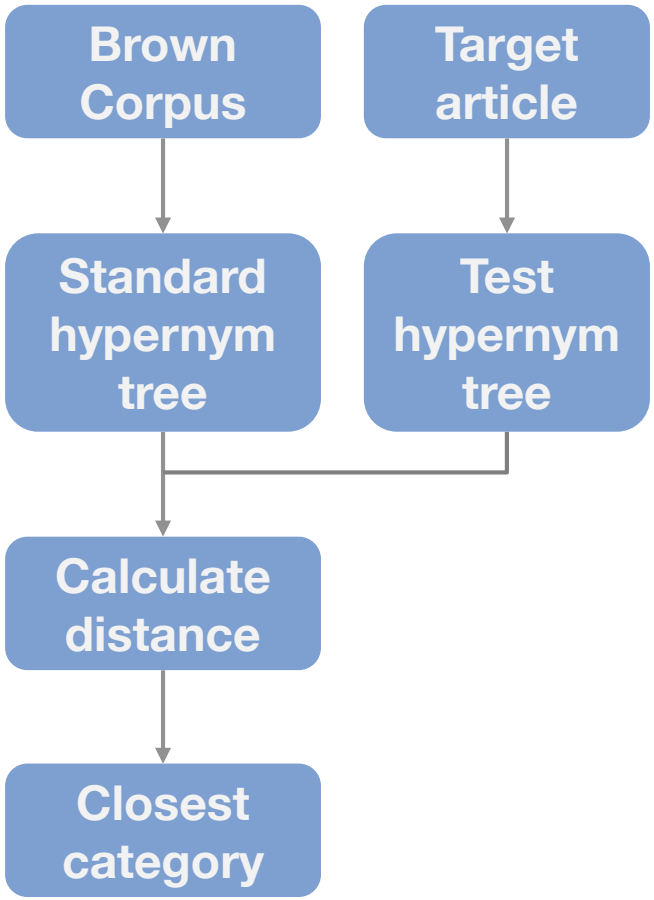
Reminder

A hyponym is a word or phrase whose semantic field is included within its **hypernym**. Since many words can have a common hypernym, **the tree structure** can be made like below figure



Popular Hypernyms : hypernyms that have relatively more hyponyms than others

Our approach : Process



1. Extract hypernym tree (V , E) from Brown Corpus
2. Calculate popular hypernym nodes from (1)
3. Make into standard FreqDict for each category
4. Calculate popular hypernym nodes for target article
5. Calculate distance (4) from (3) and find the closest category

Our approach

1. Extract hypernym tree (V, E) from Brown Corpus

File Name	Science_fiction-cm01
Number of vertices	1344
Number of edges	1221
Length of MaxPath	16
Number of Leaf Node	663

Vertices 1344

1 "know.v.01"
2 "entity.n.01"
3 "abstraction.n.06"
4 "attribute.n.02"
5 "state.n.02"
6 "condition.n.01"
7 "psychological_state.n.01"
8 "cognitive_state.n.01"
9 "consciousness.n.01"
10 "self.n.01"
.....

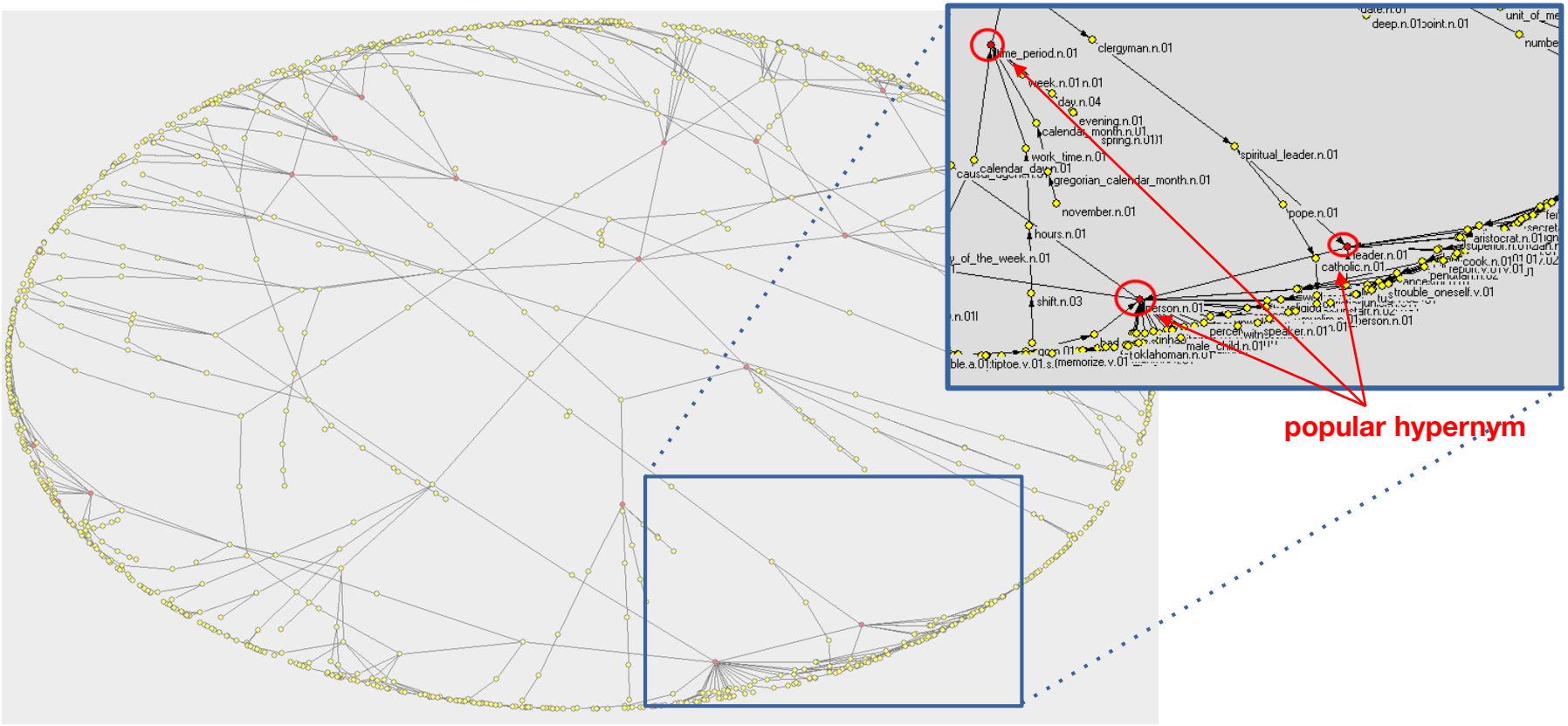
Edges 1221

3 2
4 3
5 4
6 5
7 6
8 7
9 8
10 9
11 3
12 11
....

Our approach

1. Extract hypernym tree (V, E) from Brown Corpus

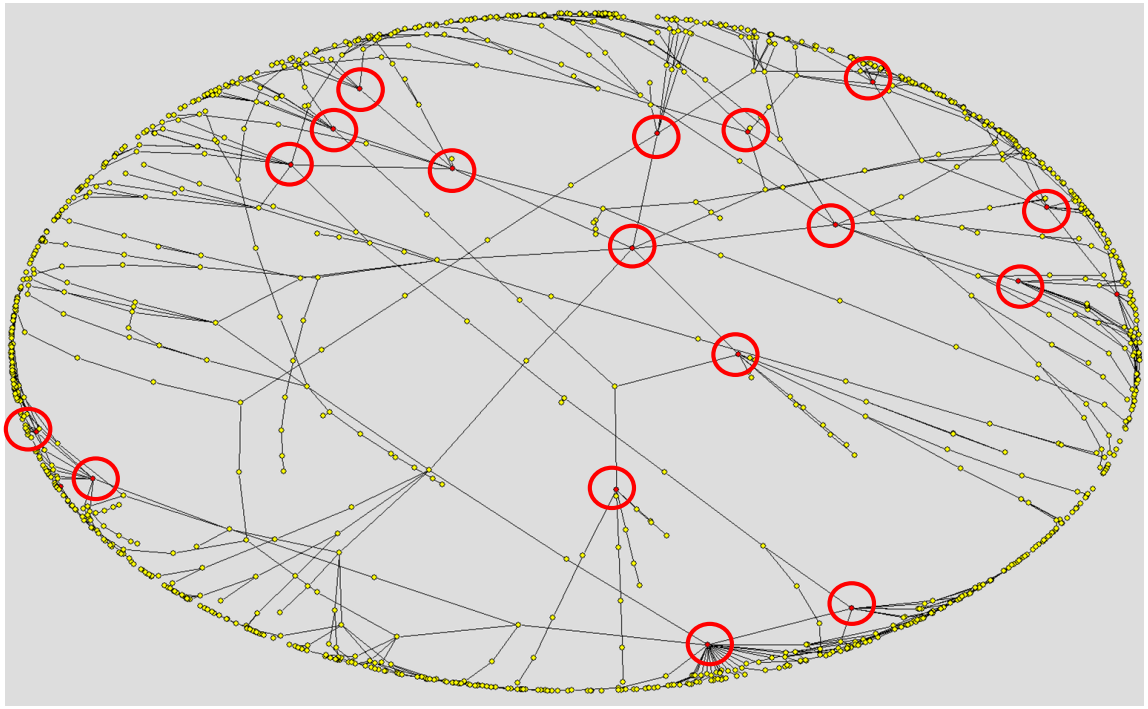
Hypernym Network drawn by pajek : science_fiction-cm01



Our approach

2. Calculate popular hypernym from (1)

Hypernym Network drawn by pajek : science_fiction-cm01



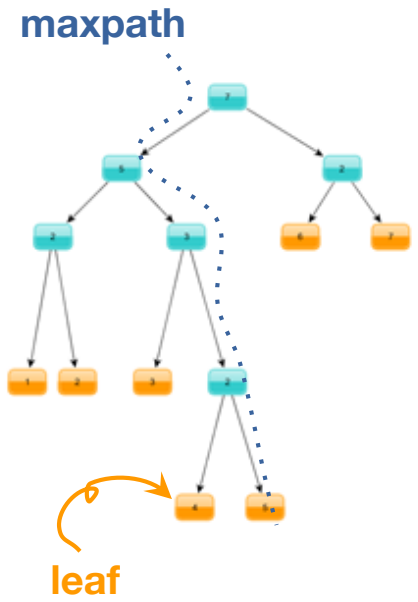
Rank	Count	Synset
0	26	person.n.01
1	11	activity.n.01
2	10	leader.n.01
3	10	property.n.02
4	9	state.n.02
5	9	artifact.n.01
6	9	act.n.02
7	9	time_period.n.01
8	9	happening.n.01
9	8	instrumentality.n.03
10	7	attribute.n.02
11	7	quality.n.01
12	7	measure.n.02
13	7	communication.n.02
14	7	message.n.02
...		

Our approach

2. Calculate popular hypernym from (1)

Statistics

category	#files	leaves	maxpaths	V	E
adventure	29	609.482759	16.37931	1243.413793	1133.241379
belles_lettres	75	649.96	15.973333	1254.706667	1137.626667
editorial	27	708.555556	16.555556	1347.185185	1233.925926
fiction	29	629.896552	15.62069	1294.172414	1193.517241
government	30	563.3	15.8	998.066667	900.633333
hobbies	36	617.694444	16.305556	1164.527778	1067.388889
humor	9	698.111111	15.444444	1411.444444	1291.444444
learned	80	554.0875	15.7375	1004.175	890.525
lore	48	645.270833	16.375	1228.104167	1116.166667
mystery	24	575.083333	15.416667	1181.708333	1081.375
news	44	669.090909	16.227273	1299.545455	1217.068182
religion	17	612.882353	16.352941	1146.235294	1032.588235
reviews	17	768.882353	16	1460.470588	1331.470588
romance	29	586.448276	15.551724	1215.62069	1115.137931
science_fiction	6	631.333333	15	1235	1110.666667



Our approach

3. Make into standard FreqDict for each category

TADA!!

net-science_fiction-cm01-info.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
#FILE:science_fiction-cm01
#Vertices 1344
#Arcs 1221
#MaxPath 16
#LeafN 663
Rank    Count   Synset
0       26      person.n.01
1       11      activity.n.01
2       10      leader.n.01
3       10      property.n.02
4       9       state.n.02
5       9       artifact.n.01
6       9       act.n.02
7       9       time_period.n.01
8       9       happening.n.01
9       8       instrumentality.n.03
10      7       attribute.n.02
11      7       quality.n.01
12      7       measure.n.02
13      7       communication.n.02
14      7       message.n.02
15      7       cognition.n.01
16      6       abstraction.n.06
17      6       group.n.01
18      6       change.v.02
19      6       device.n.01
20      6       organization.n.01
21      6       travel.v.01
22      5       physical_entity.n.01
23      5       social_group.n.01
24      5       region.n.01
25      5       structure.n.01
26      5       motion.n.06
27      5       inform.v.01
28      5       animal.n.01
29      4       condition.n.01
```

net-science_fiction-cm02-info.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
#FILE:science_fiction-cm02
#Vertices 1222
#Arcs 1101
#MaxPath 17
#LeafN 617
Rank    Count   Synset
0       31      person.n.01
1       12      activity.n.01
2       10      artifact.n.01
3       9       act.n.02
4       8       property.n.02
5       8       communication.n.02
6       7       measure.n.02
7       7       time_period.n.01
8       7       cognition.n.01
9       7       group.n.01
10      7       geographical_area.n.01
11      6       abstraction.n.06
12      6       attribute.n.02
13      6       state.n.02
14      6       inform.v.01
15      6       organization.n.01
16      5       happening.n.01
17      5       physical_entity.n.01
18      5       digit.n.01
19      5       instrumentality.n.03
20      5       communicate.v.02
21      5       feeling.n.01
22      5       content.n.05
23      5       location.n.01
24      5       change.v.02
25      5       condition.n.01
26      5       make.v.03
27      5       statement.n.01
28      5       people.n.01
29      4       relation.n.01
```

net-science_fiction-cm03-info.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
#FILE:science_fiction-cm03
#Vertices 1107
#Arcs 987
#MaxPath 15
#LeafN 564
Rank    Count   Synset
0       13      person.n.01
1       11      activity.n.01
2       10      artifact.n.01
3       9       quality.n.01
4       8       act.n.02
5       8       happening.n.01
6       7       cognition.n.01
7       7       measure.n.02
8       6       abstraction.n.06
9       6       content.n.05
10      6       attribute.n.02
11      6       communication.n.02
12      6       message.n.02
13      6       statement.n.01
14      6       concept.n.01
15      6       instrumentality.n.03
16      6       device.n.01
17      6       communicate.v.02
18      6       location.n.01
19      6       move.v.02
20      5       state.n.02
21      5       group.n.01
22      5       event.n.01
23      5       physical_entity.n.01
24      5       property.n.02
25      5       time_period.n.01
26      5       relation.n.01
27      5       body_part.n.01
28      5       group_action.n.01
29      5       organization.n.01
```

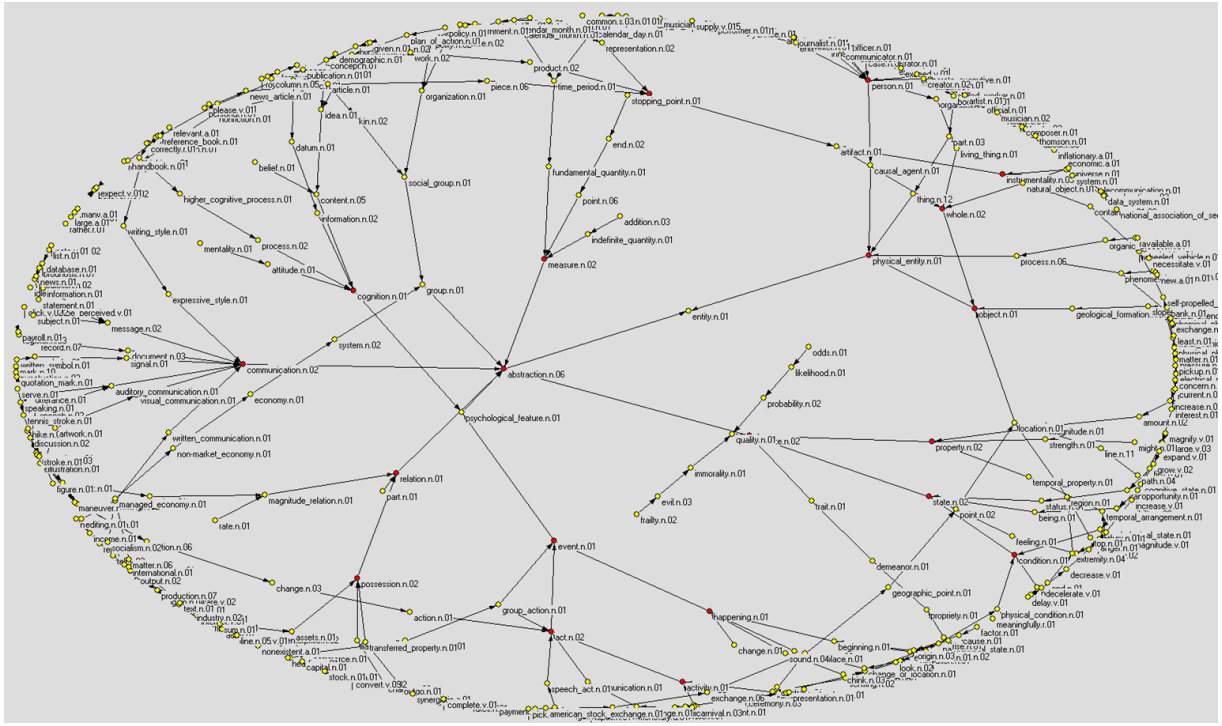
category_science_fiction_popular.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
138     person.n.01
72      activity.n.01
65      artifact.n.01
49      time_period.n.01
47      act.n.02
43      property.n.02
43      state.n.02
43      measure.n.02
41      communication.n.02
41      cognition.n.01
38      happening.n.01
38      attribute.n.02
38      quality.n.01
37      body_part.n.01
37      instrumentality.n.03
36      abstraction.n.06
34      group.n.01
33      condition.n.01
32      content.n.05
32      message.n.02
31      communicate.v.02
30      physical_entity.n.01
30      move.v.02
29      organization.n.01
29      location.n.01
```

Our approach

4. Calculate the popular hypernym of target article



Rank	Count	Synset
0	7	person.n.01
1	7	communication.n.02
2	6	abstraction.n.06
3	5	act.n.02
4	5	activity.n.01
5	4	physical_entity.n.01
6	4	cognition.n.01
7	4	attribute.n.02
8	4	state.n.02
9	4	instrumentality.n.03
10	4	happening.n.01
11	3	object.n.01
12	3	whole.n.02
13	3	event.n.01
14	3	measure.n.02
...		

Our approach

5. Calculate distance and find closest category

Standard Govenment

Rank	Count	Synset
0	532	person.n.01
1	372	activity.n.01
2	250	act.n.02
3	244	artifact.n.01
4	234	state.n.02
5	213	time_period.n.01
6	208	quality.n.01
7	194	measure.n.02
8	193	communication.n.02
9	183	large_integer.n.01
10	180	abstraction.n.06
11	178	digit.n.01
12	170	message.n.02
13	170	attribute.n.02
14	170	relation.n.01
...		

My Corpus

Rank	Count	Synset
0	7	person.n.01
1	7	communication.n.02
2	6	abstraction.n.06
3	5	act.n.02
4	5	activity.n.01
5	4	physical_entity.n.01
6	4	cognition.n.01
7	4	attribute.n.02
8	4	state.n.02
9	4	instrumentality.n.03
10	4	happening.n.01
11	3	object.n.01
12	3	whole.n.02
13	3	event.n.01
...		

Standard Romance

Rank	Count	Synset
0	685	person.n.01
1	331	artifact.n.01
2	281	activity.n.01
3	277	time_period.n.01
4	266	act.n.02
5	241	state.n.02
6	213	property.n.02
7	198	measure.n.02
8	189	body_part.n.01
9	188	instrumentality.n.03
10	180	happening.n.01
11	176	communication.n.02
12	174	abstraction.n.06
13	172	structure.n.01
14	169	quality.n.01
...		

Our approach

5. Calculate distance and find closest category

Standard Govenment

Rank	Count	Synset
0	532	per
1	372	act
2	250	a
3	244	arti
4	234	st
5	213	time_p
6	208	quality.n.01
7	194	measure.n.02
8	193	communication.n
9	183	large_integer.n.01
10	180	abstraction.n.06
11	178	digit.n.01
12	170	message.n.02
13	170	attribute.n.02
14	170	relation.n.01
...		

My Corpus

Rank	Count	Synset
6	4	cognition.n.01
7	4	attribute.n.02
8	4	state.n.02
9	4	instrumentality.n.03
10	4	happening.n.01
11	3	object.n.01
12	3	whole.n.02
13	3	event.n.01
...		

Standard Romance

Rank	Count	Synset
	685	person.n.01
	331	artifact.n.01
	281	activity.n.01
	277	time_period.n.01
	266	act.n.02
	241	state.n.02
6	213	property.n.02
7	198	measure.n.02
8	189	body_part.n.01
9	188	instrumentality.n.03
10	180	happening.n.01
11	176	communication.n.02
12	174	abstraction.n.06
13	172	structure.n.01
14	169	quality.n.01
...		

How can we calculate the distance?

Our approach

5. Calculate distance and find closest category

Our Metric
L2 Norm of index difference

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Example : two lists
L1 = [1, 4, 6], L2 = [4, 1, 6]

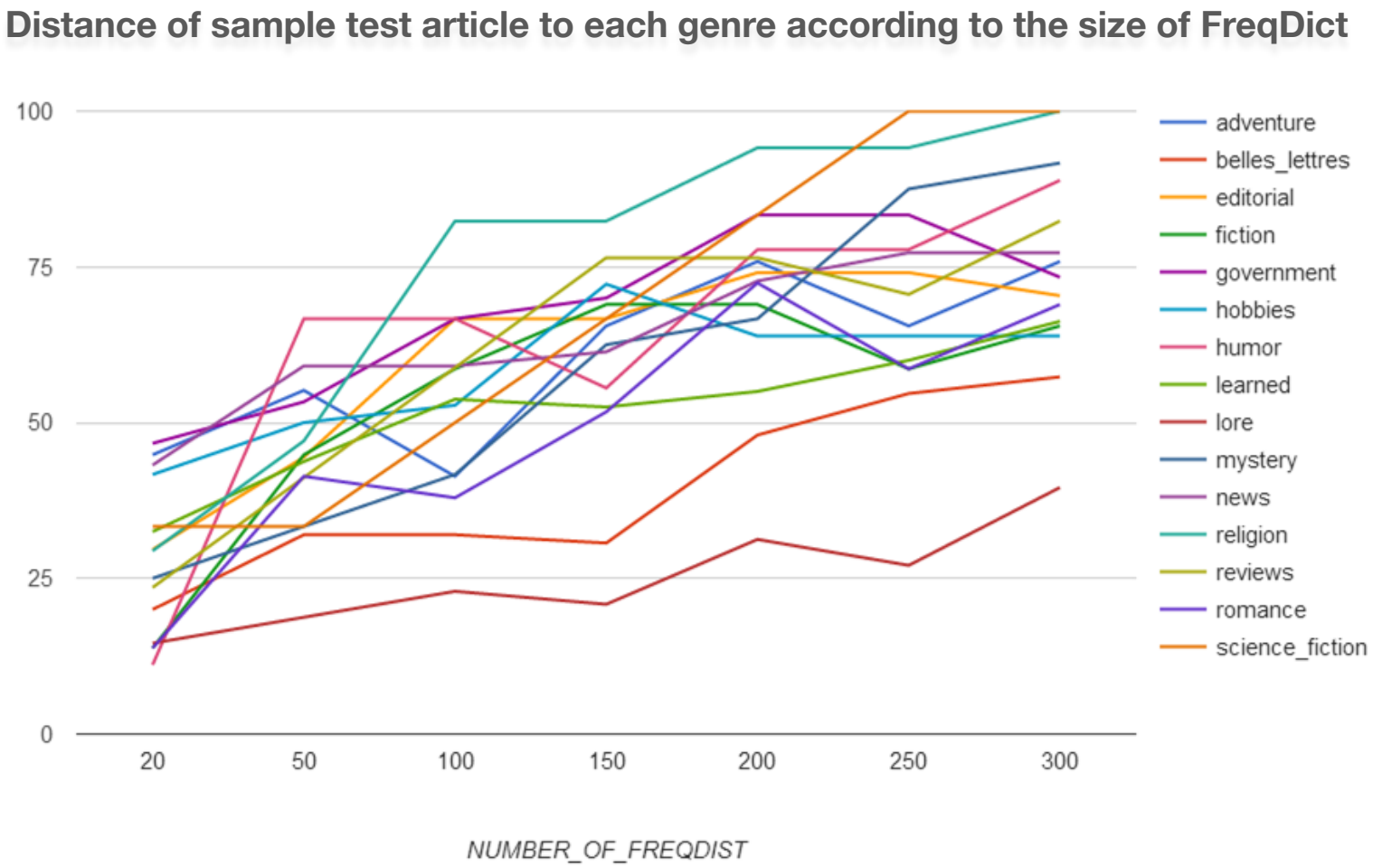
Distance of two lists
L1.index(1) == 0, L2.index(1) == 1
L1.index(4) == 1, L2.index(4) == 0
L1.index(6) == 2, L2.index(6) == 2

$$distance = \sqrt{(0-1)^2+(1-0)^2+(2-2)^2} = \sqrt{2}$$

```
def dist(L1, L2):
    d = 0
    for e in L1:
        d += (L1.index(e) - L2.index(e))**2 if e in L2 else len(L2)**2
    return math.sqrt(d)
```

Our approach

5. Calculate distance and find closest category



Results and Discussion

Results

Title of Article	1st closest	2nd closest	3rd closest
U.S. prepares plans for more troops, new base in Iraq: officials	news	editorial	belles_lettres
Bund yield hits 1 percent as stock markets halt sell-off	editorial	learned	news
At least 43 killed in Yemen clashes as parties prepare for talks	editorial	news	learned
DoubleLine's Gundlach sees odds of Fed hike by December under 5	government	editorial	learned
Pentagon bars discrimination against gays, lesbians in uniform	editorial	news	government
Apple Music faces antitrust scrutiny in NY, Connecticut	editorial	learned	government
U.S. Secretary of State Kerry tweets photo of himself in hospital	editorial	news	government
Cheap, synthetic 'flakka' dethroning cocaine on Florida drug scene	lore	government	editorial
Exclusive: Facebook earns 51 percent of ad revenue overseas	government	editorial	learned
WHO team urges South Korea to reopen schools as more close in	editorial	news	government
Convicted killer in New York prison break on third escape attempt	news	religion	fiction
Dow opens higher for first time in five days	government	news	editorial
Tokio Marine to buy HCC Insurance for $7.5 billion	editorial	news	government
Bayer sells Diabetes Care business to Panasonic Healthcare	government	learned	editorial
Spotify raises $115 million in share sale	news	government	editorial
U.S. Marine goes on trial again for killing of Iraqi civilian	editorial	belles_lettres	news
Pimco's Mather says firm expects Fed to begin raising rates in Sep	government	editorial	religion
As Greece lurches toward default, businesses hit the wall	editorial	government	lore
Putin is a 'bully,' U.S. needs to respond resolutely: Jeb Bush	editorial	news	government

Results and Discussion

Conclusion

1. Limitation of Brown corpus : only divide genre into news, romance etc

   It doesn't provide subject news in more details

2. If there is enough corpus, we need at least 300 popular hypernyms for

   reasonable results

3. Comparison of tree structure is hard! We can explore this possibility in future


Thank You for listening