Comments and answers - Assignment 5

Project Text Analysis

Leon Wetzel, Teun Buijse and Roman Terpstra

1. What is the proportion of polysemous words per Wikipedia page?

The percentages below are based on the amount of polysemous words per Wikipedia page and the amount of nouns on a page. This allows for a fair and more representative comparison, instead of using the amount of tokens. In total, 107 Wikipedia pages have been used in the application.

```
Police - 66.80%
Washington,_D.C. - 69.65%
Metropolitan_Police_Department_of_the_District_of_Columbia - 69.26%
United_States_Park_Police - 66.23%
Federal_government_of_Iraq - 73.99%
Iran - 58.71%
Islamic_Republic_News_Agency - 69.34%
Iraq - 58.71%
Tehran - 57.50%
Jalal_Talabani - 47.35%
Mahmoud_Ahmadinejad - 59.95%
Afghanistan - 57.91%
Herat_Province - 56.38%
Amanullah_Khan - 57.38%
NATO - 57.38%
Northern_Alliance - 48.79%
Zabul_Province - 50.20%
Taliban - 53.93%
World_Bank - 71.72%
Growth_Commission - 54.34%
Voice_of_America - 63.88%
Gaza_Strip - 63.07%
Gaza_City - 58.22%
Israel_Defense_Forces - 64.63%
Popular_Front_for_the_Liberation_of_Palestine - 50.63%
Gulf_Coast_of_the_United_States - 65.56%
New_Orleans - 64.03%
Barry_Wood - 64.03%
Federal_government_of_Nigeria - 66.79%
Charles_Taylor_(Liberian_politician) - 66.16%
Nigeria - 61.27%
```

```
Calabar - 52.09%
Liberia - 63.46%
Special_Court_for_Sierra_Leone - 62.59%
Sierra_Leone - 61.29%
North_Korea - 64.76%
Pyongyang - 56.62%
Bill_Clinton - 71.09%
John_Bolton - 61.76%
Musa_Qala - 61.81%
Helmand_Province - 54.87%
Associated_Press - 58.32%
NATO_Response_Force - 65.41%
British_Armed_Forces - 70.83%
United_States_Congress - 79.97%
Georgetown_University - 70.06%
Robert_Drinan - 66.31%
Massachusetts - 66.60%
United_States_House_of_Representatives - 80.34%
Pope_John_Paul_II - 62.49%
Spin_Boldak - 46.75%
Politics_of_Afghanistan - 56.95%
Hamid_Karzai - 55.31%
National_Directorate_of_Security - 41.20%
Pakistan - 58.20%
Intelligence_agency - 58.52%
Uttar_Pradesh - 52.66%
Bulandshahr_district - 43.62%
INDIA - 61.63%
Island - 73.81%
United_States - 71.29%
Guano - 56.92%
Navassa_Island - 63.16%
United_States_Coast_Guard - 73.11%
United_States_Department_of_the_Interior - 74.71%
Caribbean - 61.54%
Tokyo - 57.00%
Hong_Kong - 64.93%
Shanghai - 56.71%
Singapore - 71.27%
Taipei - 61.69%
Wellington - 63.64%
New_York_City - 62.37%
Tony_Blair - 63.85%
Al-Qaeda - 57.06%
Ayman_al-Zawahiri - 57.06%
Kingdom_of_Great_Britain - 67.04%
```

```
Videos_and_audio_recordings_of_Ayman_al-Zawahiri - 51.07%
Al_Jazeera - 63.29%
Zimbabwe - 62.44%
Democratic_Republic_of_the_Congo - 63.48%
Tanzania - 61.99%
Namibia - 62.75%
Angola - 60.78%
Mozambique - 59.86%
Botswana - 61.34%
Robert_Mugabe - 58.40%
India - 61.63%
Ministry_of_Defence_(India) - 73.79%
Chandipur,_Odisha - 42.57%
Odisha - 51.19%
Center_for_Science_in_the_Public_Interest - 65.59%
Los_Angeles - 54.23%
California - 60.36%
Video_art - 54.85%
New_media_art - 53.59%
Law_enforcement_in_Pakistan - 58.86%
Pakistan_Armed_Forces - 65.16%
Doaba - 38.03%
Hangu_District,_Pakistan - 38.03%
Peshawar - 51.09%
North_Waziristan - 46.44%
Islamabad - 54.69%
Pakistan_Army - 66.35%
Bechuanaland_Protectorate - 53.54%
Africa - 71.29%
Archipelago - 62.09%
```

2. Did all pages contain at least one polysemous word?

Based on the results in 1, we can safely assume that every page has at least one polysemous word.

3. What is the average number of senses for the polysemous words you found?

The average number of senses for polysemous words is 5.26.

4. For each number of different senses found, please list the number of words that have them (for example: 3 words showed 5 senses, 10 words showed 4 senses, and so on). Is it more or less what you would expect?

```
Senses  Occurences
11  5812
3   30724
8   10162
4   27287
7   12401
6   13776
2   40072
9   7396
10  5707
12  3603
16  1122
5   22643
15  1172
14  844
20  362
33  216
13  942
30  240
26  182
17  221
18  499
23  37
```

5. Pick five to ten cases randomly on this…

1. (Parks, Synset(‘parks.n.01’))
2. (County, Synset(‘county.n.02’))
3. (America, Synset(‘united_states.n.01’))
4. (States, Synset(‘state.n.06’))
5. (Ontario, Synset(‘ontario.n.02’))

Are they correct?

Yes, four of them are correct as they give the correct synsets.

Are they wrong?

The fourth pick seems wrong as ‘States’ indicates the US but the synset that is given indicates the word means ‘state’ and is not part of a country name.

6. Did it happen that the same word was assigned a different sense in the same Wikipedia page?

Yes, for example one time the U.S. is dexscribed as 'united_states_government.n.01' and another time as 'united_states.n.01'.

Do you think it can depend on the extention of the context you decide to use?

Yes, when more data is available the script can make better decisions when the extended text continues on the current subject, but more data can also be misleading for the script if the text continues in another subject.

7. Does what you can observe from the above correspond to what you were expecting?

Yes, this is also the case for human, when there is not a lot of context.

8. How do you think you could use disambiguation in your Wikification project?

Disambiguation can be used to make words point to the correct database more often.