Readme

Team members:
1. **Ramkishore S:** Building Distributional Thesaurus
2. **Aamir Farhan:** Evaluation of Distributional Thesaurus
3. **Atreyee:** :)

Information about DT

1. To use the complete distributional thesaurus built in python program use:
>>> from dt_20k_sim_above_3 import distributional_thesaurus
distributional_thesaurus is an list of similarity measures, it is a list of elements with the following structure:
[(<word 1>, <POS>), (<word 2>, <POS>), their similarity count]
for example look at the section titled some important results.

Files and folders

Files:

dt.py	main file to generate distributional thesaurus
dt_functions.py	functions called in dt.py, steps in creation of JoBimtext.
significance_measures.py	pmi, npmi etc
stanfordHolingOp.py	parses corpus into dependencies and applies holing operation to it.
thesauruses	contains different distributional thesauruses with different configurations

Files in folder thesauruses:

d_20k_sim_above_5.py	distributional thesaurus for all sentences with similarity measure above 5
d_20k_sim_above_3.py	distributional thesaurus for all sentences with similarity measure above 3
dt_12k_complete.py	distributional thesaurus for first 12083 sentences, can be imported into any python program as: **from dt_12k_complete import distributional_thesaurus**
dt_12k_similarity_above_9.py	similar to the above file but has only similarities with >9 context sizes
unsorted_distributional_thesaurus_first_X_sentences_with_similarity_count_greater_than_Y.txt	Thesaurus for first x sentences with similarity count > y.

unsorted_distributional_thesaurus_first_X_sentences_with_similarity_count_greater_than_Y.py	similar to first one, can be imported in similar manner, created only using first x sentences.

Files in folder parsed:

output.py	list of parses for first 2679 sentences
output2.py	list of parses for next 402 sentences
output3.py	list of parses for next 9715 sentences
output4.py	list of parses for next 3928 sentences
output5.py	list of parses for next 4164 sentences

Some interesting results:

Synonyms taken from distributional thesaurus (file: dt_20k_sim_above_5.py)

1. insects & animals:

[(u'insect', u'NN') , (u'creature', u'NN') , 8],
[(u'insect', u'NN') , (u'human', u'NN') , 6],
[(u'insect', u'NN') , (u'spider', u'NN') , 10],
[(u'insect', u'NN') , (u'snake', u'NN') , 6],
[(u'insect', u'NN') , (u'frog', u'NN') , 7],
[(u'insect', u'NN') , (u'fish', u'NN') , 9],
[(u'insect', u'NN') , (u'shark', u'NN') , 6],
[(u'insect', u'NN') , (u'bird', u'NN') , 18],
[(u'insect', u'NN') , (u'cat', u'NN') , 7],
[(u'insect', u'NN') , (u'reptile', u'NN') , 7],
[(u'insect', u'NN') , (u'beetle', u'NN') , 6],
[(u'insect', u'NN') , (u'mammal', u'NN') , 7],
[(u'insect', u'NN') , (u'species', u'NNS') , 6],
[(u'insect', u'NN') , (u'rat', u'NN') , 6],
[(u'insect', u'NN') , (u'animal', u'NN') , 9],

[(u'lizard', u'NN') , (u'snake', u'NN') , 8],
[(u'lizard', u'NN') , (u'frog', u'NN') , 7],
[(u'lizard', u'NN') , (u'gecko', u'NN') , 6],
[(u'lizard', u'NN') , (u'rat', u'NN') , 6],
[(u'lizard', u'NN') , (u'animal', u'NN') , 9],
[(u'lizard', u'NN') , (u'It', u'PRP') , 7],

[(u'chicken', u'NN') , (u'pigeon', u'NN') , 6],
[(u'chicken', u'NN') , (u'goat', u'NN') , 6],
[(u'chicken', u'NN') , (u'fish', u'NN') , 9],
[(u'chicken', u'NN') , (u'rabbit', u'NN') , 6],
[(u'chicken', u'NN') , (u'corn', u'NN') , 7],
[(u'chicken', u'NN') , (u'monkey', u'NN') , 7],
[(u'chicken', u'NN') , (u'cat', u'NN') , 7],
[(u'chicken', u'NN') , (u'horse', u'NN') , 6],

[(u'chicken', u'NN') , (u'dog', u'NN') , 9],
[(u'chicken', u'NN') , (u'sheep', u'NN') , 12],
[(u'chicken', u'NN') , (u'pig', u'NN') , 7],

[(u'breed', u'NN') , (u'breeds', u'NNS') , 9],
[(u'breed', u'NN') , (u'word', u'NN') , 6],
[(u'breed', u'NN') , (u'bird', u'NN') , 6],
[(u'breed', u'NN') , (u'cat', u'NN') , 6],
[(u'breed', u'NN') , (u'dog', u'NN') , 6],

2. days:

[(u'days', u'NNS') , (u'years', u'NNS') , 15],
[(u'days', u'NNS') , (u'minutes', u'NNS') , 6],
[(u'days', u'NNS') , (u'months', u'NNS') , 6],
[(u'days', u'NNS') , (u'types', u'NNS') , 6],

3. months:

[(u'April', u'NNP') , (u'October', u'NNP') , 6],
[(u'April', u'NNP') , (u'September', u'NNP') , 6],

4. farming:

[(u'farming', u'NN') , (u'agriculture', u'NN') , 7],
[(u'farming', u'NN') , (u'raising', u'NN') , 8],
[(u'farming', u'NN') , (u'husbandry', u'NN') , 7],
[(u'farming', u'NN') , (u'livestock', u'NN') , 7],
[(u'farming', u'NN') , (u'production', u'NN') , 9],
[(u'farming', u'NN') , (u'sheep', u'NN') , 6],
[(u'farming', u'NN') , (u'farms', u'NNS') , 8],

5. injury:

[(u'injury', u'NN') , (u'inflammation', u'NN') , 6],
[(u'injury', u'NN') , (u'disease', u'NN') , 6],
[(u'injury', u'NN') , (u'damage', u'NN') , 7],

6. offers:

[(u'offers', u'VBZ') , (u'used', u'VBD') , 6],
[(u'offers', u'VBZ') , (u'uses', u'VBZ') , 9],
[(u'offers', u'VBZ') , (u'supports', u'VBZ') , 7],
[(u'offers', u'VBZ') , (u'has', u'VBZ') , 12],
[(u'offers', u'VBZ') , (u'contains', u'VBZ') , 7],
[(u'offers', u'VBZ') , (u'includes', u'VBZ') , 9],
[(u'offers', u'VBZ') , (u'produces', u'VBZ') , 7],
[(u'offers', u'VBZ') , (u'features', u'VBZ') , 8],
[(u'offers', u'VBZ') , (u'using', u'VBG') , 6],

7. it:

[(u'It', u'PRP') , (u'creature', u'NN') , 8],
[(u'It', u'PRP') , (u'He', u'PRP') , 48],
[(u'It', u'PRP') , (u'camera', u'NN') , 11],
[(u'It', u'PRP') , (u'fish', u'NN') , 7],
[(u'It', u'PRP') , (u'cells', u'NNS') , 6],
[(u'It', u'PRP') , (u'cat', u'NN') , 7],

[(u'It', u'PRP') , (u'dog', u'NN') , 10],
[(u'It', u'PRP') , (u'worm', u'NN') , 6],
[(u'It', u'PRP') , (u'mascot', u'NN') , 9],
[(u'It', u'PRP') , (u'phone', u'NN') , 10],
[(u'It', u'PRP') , (u'liver', u'NN') , 6],
[(u'It', u'PRP') , (u'She', u'PRP') , 25],
[(u'It', u'PRP') , (u'he', u'PRP') , 17],
[(u'It', u'PRP') , (u'lake', u'NN') , 6],
[(u'It', u'PRP') , (u'version', u'NN') , 8],
[(u'It', u'PRP') , (u'They', u'PRP') , 36],
[(u'It', u'PRP') , (u'she', u'PRP') , 6],
[(u'It', u'PRP') , (u'they', u'PRP') , 11],
[(u'It', u'PRP') , (u'lizard', u'NN') , 7],
[(u'It', u'PRP') , (u'who', u'WP') , 11],
[(u'It', u'PRP') , (u'interface', u'NN') , 11],
[(u'It', u'PRP') , (u'This', u'DT') , 44],
[(u'It', u'PRP') , (u'device', u'NN') , 16],
[(u'It', u'PRP') , (u'frog', u'NN') , 8],
[(u'It', u'PRP') , (u'products', u'NNS') , 6],
[(u'It', u'PRP') , (u'article', u'NN') , 11],
[(u'It', u'PRP') , (u'larva', u'NN') , 11],
[(u'It', u'PRP') , (u'tissue', u'NN') , 6],
[(u'It', u'PRP') , (u'which', u'WDT') , 6],
[(u'It', u'PRP') , (u'system', u'NN') , 7],
[(u'It', u'PRP') , (u'that', u'WDT') , 14],
[(u'It', u'PRP') , (u'area', u'NN') , 7],
[(u'It', u'PRP') , (u'snake', u'NN') , 7],
[(u'It', u'PRP') , (u'cell', u'NN') , 8],
[(u'It', u'PRP') , (u'it', u'PRP') , 22],
[(u'It', u'PRP') , (u'screen', u'NN') , 11],
[(u'It', u'PRP') , (u'patient', u'NN') , 8],
[(u'It', u'PRP') , (u'button', u'NN') , 6],
[(u'It', u'PRP') , (u'species', u'NNS') , 12],
[(u'It', u'PRP') , (u'company', u'NN') , 8],
[(u'It', u'PRP') , (u'name', u'NN') , 6],
[(u'It', u'PRP') , (u'video', u'NN') , 6],
[(u'It', u'PRP') , (u'process', u'NN') , 6],

8. were:

[(u'were', u'VBD') , (u'are', u'VBP') , 76],
[(u'were', u'VBD') , (u'is', u'VBZ') , 67],
[(u'were', u'VBD') , (u'was', u'VBD') , 83],
[(u'were', u'VBD') , (u'being', u'VBG') , 18],
[(u'were', u'VBD') , (u'been', u'VBN') , 41],
[(u'were', u'VBD') , (u"'s", u'VBZ') , 9],
[(u'were', u'VBD') , (u'be', u'VB') , 45],

9. main:

[(u'main', u'JJ') , (u'important', u'JJ') , 10],
[(u'main', u'JJ') , (u'traditional', u'JJ') , 7],
[(u'main', u'JJ') , (u'major', u'JJ') , 6],
[(u'main', u'JJ') , (u'Other', u'JJ') , 12],
[(u'main', u'JJ') , (u'primary', u'JJ') , 8],
[(u'main', u'JJ') , (u'original', u'JJ') , 6],
[(u'main', u'JJ') , (u'new', u'JJ') , 6],
[(u'main', u'JJ') , (u'popular', u'JJ') , 6],
[(u'main', u'JJ') , (u'first', u'JJ') , 7],

10. during:

[(u'during', u'IN') , (u'In', u'IN') , 8],
[(u'during', u'IN') , (u'from', u'IN') , 7],
[(u'during', u'IN') , (u'in', u'IN') , 14],
[(u'during', u'IN') , (u'for', u'IN') , 12],
[(u'during', u'IN') , (u'to', u'TO') , 6],
[(u'during', u'IN') , (u'on', u'IN') , 8],

11. disease:

[(u'disease', u'NN') , (u'infection', u'NN') , 7],
[(u'disease', u'NN') , (u'snake', u'NN') , 4],
[(u'disease', u'NN') , (u'failure', u'NN') , 5],
[(u'disease', u'NN') , (u'necrosis', u'NNS') , 4],
[(u'disease', u'NN') , (u'death', u'NN') , 4],
[(u'disease', u'NN') , (u'organism', u'NN') , 4],
[(u'disease', u'NN') , (u'damage', u'NN') , 8],
[(u'disease', u'NN') , (u'injury', u'NN') , 6],
[(u'disease', u'NN') , (u'tumor', u'NN') , 6],
[(u'disease', u'NN') , (u'tissue', u'NN') , 5],
[(u'disease', u'NN') , (u'worm', u'NN') , 4],

12. only:

[(u'only', u'RB') , (u'However', u'RB') , 7],
[(u'only', u'RB') , (u'typically', u'RB') , 5],
[(u'only', u'RB') , (u'also', u'RB') , 23],
[(u'only', u'RB') , (u'Also', u'RB') , 4],
[(u'only', u'RB') , (u'itself', u'PRP') , 5],
[(u'only', u'RB') , (u'generally', u'RB') , 4],
[(u'only', u'RB') , (u'long', u'RB') , 4],
[(u'only', u'RB') , (u'then', u'RB') , 7],
[(u'only', u'RB') , (u'primarily', u'RB') , 5],
[(u'only', u'RB') , (u'sometimes', u'RB') , 7],
[(u'only', u'RB') , (u'there', u'RB') , 5],
[(u'only', u'RB') , (u'still', u'RB') , 4],
[(u'only', u'RB') , (u'here', u'RB') , 5],
[(u'only', u'RB') , (u'Only', u'RB') , 4],
[(u'only', u'RB') , (u'now', u'RB') , 6],
[(u'only', u'RB') , (u'occasionally', u'RB') , 5],
[(u'only', u'RB') , (u'commonly', u'RB') , 4],
[(u'only', u'RB') , (u'currently', u'RB') , 4],
[(u'only', u'RB') , (u'when', u'WRB') , 6],
[(u'only', u'RB') , (u'often', u'RB') , 5],
[(u'only', u'RB') , (u'usually', u'RB') , 10],
[(u'only', u'RB') , (u'always', u'RB') , 4],
[(u'only', u'RB') , (u'mainly', u'RB') , 5],