ukWaC word2vec Workflow

C, OpenMP based Program

ukWaC → **Run ./wacky /path/to/ukwac/files** ← dictionary size: 50,000

- **freq.txt**
- **dictionary.txt**
- **subjects.txt**
- **integers_<filename>_<thread>.txt**
- **total_count.txt**
- **unk_count.txt**
- **size_<filename>_<thread>.txt**
- **words_<filename>_<thread>.txt**

words_<filename>_<thread>.txt → combine word files → all words concatenated

Original C based Google Program

all words concatenated → run word2vec ←
```
cbow: 1
size: 200
window: 8
negative: 25
hs: 0
sample: 1e-4
threads: 20
iter: 15
```

run word2vec → vectors.bin

Local Python 3 Script

vectors.bin → python3 verbs.py /path/to/wacky/output/files -b

dictionary.txt → python3 verbs.py /path/to/wacky/output/files -b
subjects.txt → python3 verbs.py /path/to/wacky/output/files -b

python3 verbs.py → cosine distances → **End**