NLP PROJECT

AN EXPERIMENT FOR WORD NET ENHANCEMENT

T.S.V.MANEESH(201201183)

B.CHAITANYA(201101163)

1. INTRODUCTION

WordNet is a lexical database for a language. It groups words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.

Some reasonably high frequency words do not have synsets in the wordnet database. So, we suggest a synset for such words by taking most similar words using cosines.

2. INPUT

- Big corpus: - This data is used to enhance the wordnet. It contains 4.3 million lines of Hindi sentences in WX format.
- Wordnet database: - It contains 38 thousand synsets in UTF format.

3. OUTPUT

- Suggest synset for a word: - If a word does not already have a synset, a synset of 20 words is suggested for the word.
- Suggesting a word to a synset: - If a word is not present in any synset then it is suggested to the best suitable synset.

4. PROCEDURE

i. Wordnet data is in UTF format , convert it into WX using UTF to WX tool

ii. Run word2vec tool on raw data and build model.

iii. Get vocabulary and their respective counts from the corpus.

iv. Parse the wordnet database and extract synsets.

v. For reasonably high frequency words which o not have a synset in the wordnet database , extract the top 20 similar words using cosine similarity and suggest synset.

vi. For reasonably high frequency words which are not present in any synset, suggest the best suitable synset.

5. ASSUMPTIONS

- For step (v. in Procedure), we assumed ‘reasonably high frequency‘ to be in the range 100,000 – 43.
- For step (vi. in Procedure), we assumed ‘reasonably high frequency‘ to be in the range 100,000 – 1000.
- To suggest the best suitable synset, we compared the word with all entries in the synset and took the synset with highest average cosine similarity.

6. OSERVATIONS

- A synset should have words with similar meaning but because we suggested synset of length 20, some words with opposite meanings and with little similarities are also present in the synset.
- Suggested synset also contains adjectives if the word is a noun.

7. RESULTS

Sample output for step v.

बारात शुग्गेस्टेड् स्य्न्सेट्:

[बाजा,बरात,शहनाई,खंडाला,महाबलेश्वर,मेंहदी,बाराती,सेरमनी,नगाड़ा,ध्रुपद,बराती,वधु,इकतारा,दुल्हनिया,ओंकारा,बैंडबाजे,दूल्हन,पियानो,पेंटी,ढोल]

अविवाहित फ्रेसेन्ट् इन् तोर्ड्नेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्:
[अनब्याहा,कुँआरा,कुँवारा,क्वाँरा,कुंवार,कँवारा,क्वारा,बिनब्याहा,अनूढ़,अपरिणीत,ग़ैर_शादीशुदा,गैर_शादीशुदा]

प्रतिलिपि शुग्गेस्टेड् स्य्न्सेट्:
[प्रपत्र,पाक्षिक,प्रेषक,हस्तलिखित,साप्,ग्रंथालय,जोखा,अभिलेख,सचित्र,रपट,राजपत्र,नोट्स,गाइडलाइंस,डोमिनियन,प्रोटोकोल,छपवा,नियमावली,आलेख,परिपत्र,प्रतिवेदन]

असमी शुग्गेस्टेड् स्य्न्सेट्:
[कोंकणी,प्राकृत,नागरी,कथाकार,असमिया,संथाली,गोंडी,घुमक्कड,अवधी,मैथिली,पाठांतर,उड़िया,अपभ्रंश,बांग्,रुपान्तर,छत्तीसगढ़ी,निमाड़ी,पंचतंत्र,ऊर्दू,विश्वकोश]

मेखला फ्रेसेन्ट् इन् तोर्ड्नेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: [कफ़नी,कफनी,अलफी]

डौल शुग्गेस्टेड् स्य्न्सेट्:
[काठी,छरहरा,मिस्री,नुमा,गठीला,सैंटीमीटर,लम्बाई,टाप,पूर्णांक,मेहराब,डीलडौल,पैंसिल,मरकरी,थुलथुल,शंकु,ठिगना,नैक,ट्वाइन,पाजेब,गतिया]

बाराती शुग्गेस्टेड् स्य्न्सेट्:
[राहगीर,ट्रैक्टरट्रॉली,कांवरियों,क्वालिस,बराती,टैंपो,तीर्थयात्री,मनचला,बोलेरो,कांवड़ियों,मांबेटी,कांवड़िए,प्रेमप्रसंग,ट्राले,श्रद्धालु,पीछेपीछे,लूटेरों,कांवडिय़ों,टेंपो,युवकयुवती]

दुधारू फ्रेसेन्ट् इन् तोर्ड्नेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: [दुधार,दुधैल,दुधारी,दुधैली]

आख़िरकार शुग्गेस्टेड् स्य्न्सेट्:
[आखिरकार,यूलियांती,ख़ुद,अपनाअपना,एनेल्का,अंतत,यिप,डूंगा,रोबेन,कार्लोविच,सोडरलिंग,रेजाई,विट्टेक,मेज़बान,नीदरलैंड्स,उरुग्वे,गोलरहित,सरीना,मोल्स,पराग्वे]

उपकेंद्र शुग्गेस्टेड् स्य्न्सेट्:
[रतनपुर,हरसूद,विश्रामगृह,शिवपुर,नरसिंहगढ़,पीथमपुर,कन्नौद,ब्यावरा,आरौन,खुजनेर,बेरछा,बाजना,अरेर,निम्बाहेड़ा,विद्यानगर,लालपुर,साँवेर,रानापुर,बदरवास,देपालपुर]

असम फ्रेसेन्ट् इन् तोर्ड्नेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: []

नेतृत्वकर्ता शुग्गेस्टेड् स्य्न्सेट्:
[चिंतक,शीर्षस्थ,टिप्पणीकार,एसोसियेशन,हिंदुत्ववादी,भारतीयअमेरिकी,मार्गदर्शक,अगुवा,सामजिक,वामपंथ,राष्ट,अस्मिता,हर्मन,मनोविज्ञानी,यूएनडीपी,अम्मान,राष्ट्रिय,ख्यातिप्राप्त,कम्युनिष्ट,सदस्या]

फ्रैंड्स शुग्गेस्टेड् स्यन्सेट्:
[जोक्स,एड्रेस,प्रोडक्ट,गिफ्ट,यूजर,डायरी,जेनरेशन,होस्टिंग,आवेदक,यूजर्स,ऑफिशियल,आईडी,मैप्स,मॉर्निंग,जीवनसाथी,जीमेल,धारक,कुरियर,लैटर,ग्राहक]

लिखित फ्रेसेन्ट् इन् तोई्नेट् अल्रेअड्य् तिट्ह् स्यन्सेट्: [लिपिबद्ध,अंकित,लिखा,लिखा_हुआ,मकतूब]

270 शुग्गेस्टेड् स्यन्सेट्:
[190,160,260,195,155,265,130,275,135,235,470,180,140,120,165,280,185,399,145,240]

271 शुग्गेस्टेड् स्यन्सेट्:
[186,246,299,382,208,256,157,156,141,293,402,171,136,138,143,263,275,277,273,760]

272 शुग्गेस्टेड् स्यन्सेट्:
[182,326,271,288,545,403,543,195,294,दोतिहाई,273,119,126,230,383,सीटे,143,208,223,162]

कैरोलिना शुग्गेस्टेड् स्यन्सेट्:
[इंडियाना,मिसौरी,वर्जीनिया,नेवादा,फ़्लोरिडा,ओहायो,पेंसिलवेनिया,कोलंबिया,ओहियो,कोलोराडो,टेक्सास,मेरीलैंड,केरोलिना,मिनेसोटा,कैलीफोर्निया,पापुआ,अलाबामा,कैलिफोर्निया,मैरीलैंड,नेब्रास्का]

आंतकी शुग्गेस्टेड् स्यन्सेट्:
[आतंकी,आंतकवादी,आतंकवादी,मास्टरमाइंड,लश्करएतैयबा,लश्कर,2611,साजिशकर्ता,लश्करएतैबा,चरमपंथी,एलईटी,हूजी,आईएम,जैशएमोहम्मद,हिजबुल,लश्करए,आंतकियों,अलशबाब,लश्करएतोएबा,जमातउददावा]

275 शुग्गेस्टेड् स्यन्सेट्:
[271,185,246,208,760,155,135,270,352,265,384,160,260,186,290,195,255,159,278,435]

276 शुग्गेस्टेड् स्यन्सेट्:
[267,338,455,402,229,459,760,352,384,362,197,332,299,484,252,172,445,149,387,425]

नागराज शुग्गेस्टेड् स्यन्सेट्:
[बसंता,गरूड़,गोह,मूस,वासुकि,सूर्यदेव,कासा,असुर,सापं,श्रृंगी,कैक्टस,आद्य,पेरुमाल,काकड़,कापालिक,नीलकंठ,सागवान,सुब्रहमण्,रामानुज,शालिग्राम]

278 शुग्गेस्टेड् स्यन्सेट्:
[255,209,238,185,265,246,202,384,285,297,232,352,760,172,142,236,208,346,245,385]

सौभाग्य फ्रेसेन्ट् इन् तोई्नेट् अल्रेअड्य् तिट्ह् स्यन्सेट्: [खुशक़िस्मती,खुशनसीबी,सद्भाग्य,सआदत]

नादिया फ्रेसेन्ट् इन् तोई्नेट् अल्रेअड्य् तिट्ह् स्यन्सेट्: [नंदी,नन्दी,नादिया_बैल,नंदी_बैल,नन्दी_बैल]

स्वात फ्रेसेन्ट् इन् तोर्ईनेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: []

जोएल शुग्गेस्टेड् स्य्न्सेट्:
[गार्नर,मैल्कम,चैपमैन,जियान,गिटारवादक,कॉलिन,जैफ,जैमी,केरेन,गेविन,अंडरवुड,बेवन,गूच,मॉरिसन,जोए,कैरोल,पैटिनसन,शेफर्ड,जेफ़,जेफ]

पोलक फ्रेसेन्ट् इन् तोर्ईनेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: []

आपराधिक फ्रेसेन्ट् इन् तोर्ईनेट् अल्रेअड्य् तिट्ह् स्य्न्सेट्: []

स्वाती शुग्गेस्टेड् स्य्न्सेट्:
[चंपक्का,संती,भनिता,तिम्मम्मा,बकुली,सिध्दार्थ,चीनिवास,किशनसिंह,चंद्रक्कारन,धरणी,किट्टी,अमफू,वसंतराव,कोच्चम्मा,तरुलता,अप्पु,दीपे,लीखे,अक्कम्मा,फेदी]

चैकअप शुग्गेस्टेड् स्य्न्सेट्:
[चेकअप,ईलाज,आप्रेशन,अस्पातल,एसकेआईएमएस,एलएनजेपी,अल्ट्रासाऊंड,जीएमसी,सीएमसी,ट्रांसप्लांट,एडमिट,ईसीजी,डीएमसी,डाक्टरी,अस्तपाल,ईएनटी,मैडीकल,एमरजैंसी,पैथोलॉजी,मैडिकल]

आंतक शुग्गेस्टेड् स्य्न्सेट्:
[अराजकता,जनाक्रोश,विद्वेष,अलगाववाद,प्रतिशोध,गुंडागर्दी,नापाक,कट्टरपंथ,विघटनकारी,बर्बरता,अशांति,संप्रदायवाद,कत्लेआम,भ्रष्टचार,खौफ,दहशतगर्द,संवेदनहीनता,तानाबाना,कुचक्र,असहिष्णुता]

भारतीयअमेरिकी शुग्गेस्टेड् स्य्न्सेट्:
[ह्यूग,ऑस्ट्रियाई,हैंपशायर,ब्रूनो,वास्तुकार,चैपमैन,मेसाचुसेट्स,जूड,क्रिस्टोफ,गैब्रिएल,नेतृत्वकर्ता,एसोसियेशन,कनाडियाई,प्रेस्टन,ओपराह,एंटोनियो,फेंडर,पियर,क्लारा,ब्रुस]

कुरुप शुग्गेस्टेड् स्य्न्सेट्:
[सत्यकाम,वीरेश्वर,हरिभाई,मनुभाई,रमणी,मुनी,हेग्गड़ेजी,भगवन्,कविराज,सिध्दार्थ,गुरुस्वामी,गोशाल,धाई,शक्तिदेव,चंद्रस्वामी,रंभाजी,सुनयनी,श्रृंगभुज,मनसुख,ब्राह्मण]

खगोल शुग्गेस्टेड् स्य्न्सेट्:
[भौतिकी,शास्त्र,वैज्ञानिक,विज्ञान,मानविकी,भूविज्ञान,प्राणि,समाजशास्त्र,विज्ञानी,मनोविज्ञान,भूवैज्ञानिक,अर्थशास्त्र,संकाय,गणित,भूभौतिकी,भूगोल,न्यूक्लीय,हस्तरेखा,खगोलविद्,भूगर्भ]

इकलौता शुग्गेस्टेड् स्य्न्सेट्:
[इकलौती,भाई,आरव,नौरीन,कुँवारा,चहेता,एकमात्र,क़रीबी,नाती,पुत्र,यासमीन,तीनो,स्टैलोन,बेटा,फ़िरोज़,भतीजा,अय्याश,कमाऊ,भरोसेमंद,चचेरा]

Sample output for step vi.

पेपर श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [समाचारपत्र,समाचार-पत्र,समाचार_पत्र,अख़बार,अखबार,पेपर,न्यूज़_पेपर,न्यूज_पेपर,न्यूज़पेपर,न्यूजपेपर]

संतुलन भेलोन्ग्स् टो स्यन्सेट्: [संतुलन,]

दोस्ती भेलोन्ग्स् टो स्यन्सेट्: [दोस्ती,यारी,मित्रता,मैत्री,याराना,बंधुता,मिताई,दोस्तदारी,सौहार्द,सौहार्द्य,मेल,उलफत,उलफ़त,इखलास,मुआफ़िक़त,मुआफिकत,मुआफकत,रफ़ाकत,रफाकत,इख़्तिलात,इख्तिलात,इठाई,इष्टता,ईठि,वास्ता]

कमिटी श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [समिति,कमेटी,कमिटी,पैनल,पेनल,कमीशन,कमिशन]

कलाकार भेलोन्ग्स् टो स्यन्सेट्: [कलाकार,फ़नकार,फनकार,कलाकर्मी,हुनरमंद,हुनरमन्द,आर्टिस्ट]

मीडिया श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [रपटा,रपट,रपट्टा]

हट श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [ठीक,ठीक_से,अच्छी_तरह,अच्छी_तरह_से,सुचारु_रूप_से]

स्वराज श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [स्वराज्य,स्वराज,सुराज]

क्रियान्वयन श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [कार्यान्वयन,]

गौतम श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [गौतम_बुद्ध,बुद्ध,गौतम,भगवान_बुद्ध,तथागत,विश्वबोध,बुद्धदेव,सिद्धार्थ,विश्वंतर,विश्वन्तर,वीतराग,धर्मकाय,धर्मकेतु,महाश्रमण,दम,सरल,करुण]

सीआरपीएफ श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [बल,]

रोमांटिक श्होउल्ड् बेलोन्ग् टो स्यन्सेट्: [काक्टेल,काकटेल,कॉक्टेल,कॉकटेल]

विश्वविद्यालय भेलोन्ग्स् टो स्यन्सेट्: [विश्वविद्यालय,विद्यापीठ,यूनिवर्सिटी,युनिवर्सिटी]

झगड़ा भेलोन्ग्स् टो स्यन्सेट्: [झगड़ा,कलह,विवाद,बखेड़ा,लड़ाई,लफड़ा,लड़ाई-झगड़ा,झगड़ा-लड़ाई,फसाद,फ़साद,ख़ुराफ़ात,खुराफात,झंझट,चकल्लस,फुतूर,फ़ुतूर,फतूर,फ़तूर,लडाई-भिड़ाई,भिड़Mत,भिड़न्त,टंटा,खटराग,झड़प,अड़प-

झड़प,लोचा,रार,राड़,षट्राग,विग्रह,चकरबा,अनुशय,अपड़ाव,अभिग्रह,अभेरा,रैसा,रैहर,अरवाह,अवडेर,निज़ा,निजा,निज़ाअ,निजाअ,अवरेब,ईति]

उपमुख्यमंत्री भेलोन्ग्स् टो स्य्न्सेट्: [उपमुख्यमंत्री,उप_मुख्यमंत्री,उप-मुख्यमंत्री,डिप्टी_सीएम,डिप्टी_सी_एम]

राजस्थान भेलोन्ग्स् टो स्य्न्सेट्: [राजस्थान,राजपूताना]

मेहमान श्होउल्ड् बेलोन्ग् टो स्य्न्सेट्: [कीवी,]

दोस्त श्होउल्ड् बेलोन्ग् टो स्य्न्सेट्: [मित्र,मित्र_देव,मित्र_देवता]

रूपया श्होउल्ड् बेलोन्ग् टो स्य्न्सेट्: [पाकिस्तानी_रुपया,पाकिस्तानी_रुपिया,रुपया,रुपिया]

आरती भेलोन्ग्स् टो स्य्न्सेट्: [आरती,निराजन]

गंभीर श्होउल्ड् बेलोन्ग् टो स्य्न्सेट्: [नाजुक,नाज़ुक]

नवंबर श्होउल्ड् बेलोन्ग् टो स्य्न्सेट्: [जुलाई,]