

# **NLP PROJECT**

## **AN EXPERIMENT FOR WORD NET ENHANCEMENT**

**Sandeep Sricharan  
Mukku(201202157)**

## **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 do 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. OBSERVATIONS**

- 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 स्यन्सेट्:

[बाजा  
,बारात,शहनाई,खंडाला  
,महाबलेश्वर,मेंहदी

,बाराती  
,सेरमनी,नगाड़ा,ध्रुपद,बराती,वधु,इकतारा,दुल्हनिया,ओंकारा,बैंडबाजे,दूल्हन,पियानो,पेंटी,ढोल]

अविवाहित फ्रेसेन्ट इन् तोर्डनेट् अल्रेअडच् तिट्ह् स्यन्सेट्:

[अनब्याहा

,कुँआरा

,कुँवारा,क्वाँरा,कुँवार,कँवारा,क्वारा,बिनब्याहा,अनूढ़,अपरिणीत,गैर\_शादीशुदा,गैर\_शादीशुदा]

प्रतिलिपि शुग्गेस्टेड् स्यन्सेट्:

[प्रपत्र,पाक्षिक,प्रेषक,हस्तलिखित,साप्

,ग्रंथालय,जोखा

,अभिलेख,सचित्र,रपट,राजपत्र,नोट्स,गाइडलाइंस,डोमिनियन,प्रोटोकोल,छपवा

,नियमावली,आलेख,परिपत्र,प्रतिवेदन]

असमी शुग्गेस्टेड् स्यन्सेट्:

[कोंकणी

,प्राकृत,नागरी

,कथाकार,असमिया

,संथाली

,गोंडी

,घुमक्कड,अवधी

,मैथिली,पाठांतर,उड़िया,अपभ्रंश,बांग्,रूपान्तर,छत्तीसगढ़ी,निमाड़ी,पंचतंत्र,ऊर्दू,विश्वकोश]

मेखला फ्रेसेन्ट इन् तोर्डनेट् अल्रेअडच् तिट्ह् स्यन्सेट्: [कफनी,कफनी,अलफी]

डौल शुग्गेस्टेड् स्यन्सेट्:

[काठी

,छरहरा

,मिस्री

,नुमा

,गठीला

,सैंटीमीटर,लम्बाई,टाप,पूर्णांक,मेहराब,डीलडौल,पैसिल,मरकरी

,थुलथुल,शंकु,ठिगना,नैक,ट्वाइन,पाजेब,गतिया]

बाराती शुग्गेस्टेड् स्यन्सेट्:

[राहगीर,ट्रैक्टरट्रॉली

,कांवरियों

,क्वालिस,बराती

,टैंपो

,तीर्थयात्री

,मनचला

,बोलेरो

,कांवड़ियों

,मांबेटी,कांवड़िए,प्रेमप्रसंग,ट्राले,श्रद्धालु,पीछेपीछे,लूटेरों,कांवड़ियों,टैंपो,युवकयुवती]

दुधारू फ्रेसेन्ट इन् तोर्डनेट् अल्रेअडच् तिट्ह् स्यन्सेट्: [दुधार,दुधैल,दुधारी,दुधैली]

आखिरकार शुग्गेस्टेड् स्यन्सेट्:

[आखिरकार,यूलियांती

,खुद,अपनाअपना

,एनेल्का  
,अंतत,यिप,डूंगा  
,रोबेन,कालोविच,सोडरलिंग,रेजाई,विट्टेक,मेज़बान,नीदरलैंड्स,उरुगे  
,गोलरहित,सरीना,मोल्स,पराग्वे]

उपकेंद्र शुगोस्टेड् स्यन्सेट्:

[रतनपुर,हरसूद,विश्रामगृह,शिवपुर,नरसिंहगढ़  
,पीथमपुर,कन्नौद,ब्यावरा  
,आरौन,खुजेर,वेरछा  
,बाजना,अरेर,निम्बाहेड़ा,विद्यानगर,लालपुर,साँवेर,रानापुर,बदरवास,देपालपुर]

असम फ़रेसेन्ट् इन् तोर्डनेट् अलरेअडच् तिट्ह् स्यन्सेट्: []

नेतृत्वकर्ता शुगोस्टेड् स्यन्सेट्:

[चिंतक,शीर्षस्थ,टिप्पणीकार,एसोसियेशन,हिंदुत्ववादी  
,भारतीयअमेरिकी  
,मार्गदर्शक,अगुवा  
,सामाजिक,वामपंथ,राष्ट्र,अस्मिता  
,हर्मन,मनोविज्ञानी,यूएनडीपी,अम्मान,राष्ट्रिय,ख्यातिप्राप्त,कम्युनिष्ट,सदस्या]

फ़रैंड्स शुगोस्टेड् स्यन्सेट्:

[जोक्स,एंडरेस,प्रोडक्ट,गिफ्ट,यूजर,डायरी  
,जेनरेशन,होस्टिंग,आवेदक,यूजर्स,ऑफिशियल,आईडी  
,मैप्स,मॉनिंग,जीवनसाथी,जीमेल,धारक,कुरियर,लैटर,ग्राहक]

लिखित फ़रेसेन्ट् इन् तोर्डनेट् अलरेअडच् तिट्ह् स्यन्सेट्:

[लिपिबद्ध,अंकित,लिखा,लिखा\_हुआ,मकतूब]

270 शुगोस्टेड् स्यन्सेट्:

[190,160,260,195,155,265,130,275,135,235,470,180,140,120,165,280,18  
5,399,145,240]

271 शुगोस्टेड् स्यन्सेट्:

[186,246,299,382,208,256,157,156,141,293,402,171,136,138,143,263,27  
5,277,273,760]

272 शुगोस्टेड् स्यन्सेट्:

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

कैरोलिना शुगोस्टेड् 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:

[271,185,246,208,760,155,135,270,352,265,384,160,260,186,290,195,25  
5,159,278,435]

276 शुग्गेस्टेड् स्यन्सेट्:

[267,338,455,402,229,459,760,352,384,362,197,332,299,484,252,172,44  
5,149,387,425]

नागराज शुग्गेस्टेड् स्यन्सेट्:

[बसंता  
,गरूड  
,गोह,मूस,वासुकि  
,सूर्यदेव,कासा  
,असुर,साप  
,शरृंगी,कैक्टस,आद्य,पेरुमाल,काकड,कापालिक,नीलकंठ,सागवान,सुब्रह्मण्,रामानुज,शालिग्राम]

278 शुग्गेस्टेड् स्यन्सेट्:

[255,209,238,185,265,246,202,384,285,297,232,352,760,172,142,236,20  
8,346,245,385]

सौभाग्य फ्रेसेन्ट् इन् तोर्डनेट् अल्रेअडच् तिट्हु स्यन्सेट्:

[खुशकिस्मती,खुशनसीबी,सद्भाग्य,सआदत]

नादिया फ्रेसेन्ट् इन् तोर्डनेट् अल्रेअडच् तिट्हु स्यन्सेट्:

[नंदी,नन्दी,नादिया\_बैल,नंदी\_बैल,नन्दी\_बैल]

स्वात फ्रेसेन्ट् इन् तोर्डनेट् अल्रेअडच् तिट्हु स्यन्सेट्: []

जोएल शुग्गेस्टेड् स्यन्सेट्:

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

पोलक फ्रेसेन्ट् इन् तोर्डनेट् अल्रेअडच् तिट्हु स्यन्सेट्: []

आपराधिक फ्रेसेन्ट् इन् तोर्डनेट् अल्रेअडच् तिट्हु स्यन्सेट्: []

स्वाती शुग्गेस्टेड् स्यन्सेट्:

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

चैकअप शुग्गेस्टेड् स्यन्सेट्:

[चेकअप,ईलाज,आप्रेशन,अस्पातल,एसकेआईएमएस,एलएनजेपी  
,अल्ट्रासाऊंड,जीएमसी  
,सीएमसी  
,ट्रांसप्लांट,एडमिट,ईसीजी  
,डीएमसी,डाक्टरी,अस्तपाल,ईएनटी,मैडीकल,एमरजेंसी,पैथोलॉजी,मैडिकल]

आंतक शुग्गेस्टेड् स्यन्सेट्:

[अराजकता  
,जनाक्रोश,विद्वेष,अलगाववाद,प्रतिशोध,गुंडागर्दी  
,नापाक,कट्टरपंथ,विघटनकारी  
,बर्बरता  
,अशांति  
,संप्रदायवाद,कल्लेआम,भ्रष्टचार,खौफ,दहशतगर्द,संवेदनहीनता,तानाबाना,कुचक्र,असहिष्णुता]

भारतीयअमेरिकी शुग्गेस्टेड् स्यन्सेट्:

[ह्यूग,ऑस्ट्रियाई,हैंपशायर,ब्रूनो  
,वास्तुकार,चैपमैन,मेसाचुसेट्स,जूड,क्रिस्टोफ,गैब्रिएल,नेतृत्वकर्ता  
,एसोसियेशन,कनाडियाई,प्रेस्टन,ओपराह,एंटोनियो,फेंडर,पियर,क्लारा,ब्रूस]

कुरुप शुग्गेस्टेड् स्यन्सेट्:

[सत्यकाम,वीरेश्वर,हरिभाई,मनुभाई,रमणी  
,मुनी  
,हेगगड़ेजी  
,भगवन्  
,कविराज,सिध्दार्थ,गुरुस्वामी  
,गोशाल,धाई,शक्तिदेव,चंद्रस्वामी,रंभाजी,सुनयनी,श्रृंगभुज,मनसुख,ब्राह्मराण]

खगोल शुग्गेस्टेड् स्यन्सेट्:

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

इकलौता शुग्गेस्टेड् स्यन्सेट्:

[इकलौती  
,भाई,आरव,नौरीन,कुंवारा  
,चहेता]

,एकमात्र,करीबी  
,नाती,पुत्र,यासमीन,तीनो,स्टैलोन,बेटा,फ़िरोज़,भतीजा,अय्याश,कमाऊ,भरोसेमंद,चचेरा]

## Sample output for step vi.

पेपर शहोउल्ड् बेलोन्ग् टो स्यन्सेट्: [समाचारपत्र,समाचार-  
पत्र,समाचार\_पत्र,अखबार,अखबार,पेपर,न्यूज़\_पेपर,न्यूज़\_पेपर,न्यूज़पेपर,न्यूज़पेपर]

संतुलन भेलोन्स् टो स्यन्सेट्: [संतुलन,]

दोस्ती भेलोन्स् टो स्यन्सेट्:

[दोस्ती  
,यारी  
,मित्रता  
,मैत्री  
,याराना  
,बंधुता  
,मिताई,दोस्तदारी  
,सौहार्द,सौहार्दय,मेल,उलफत,उलफ़त,इखलास,मुआफ़िकत,मुआफ़िकत,मुआफ़कत,रफ़ाकत,रफ़ाकत,  
इख़्तिलात,इख़्तिलात,इठाई,इष्टता,ईठि,वास्ता]

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

कलाकार भेलोन्स् टो स्यन्सेट्:

[कलाकार,फ़नकार,फनकार,कलाकर्मी,हुनरमंद,हुनरमन्द,आर्टिस्ट]

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

हट शहोउल्ड् बेलोन्ग् टो स्यन्सेट्:

[ठीक,ठीक\_से,अच्छी\_तरह,अच्छी\_तरह\_से,सुचारु\_रूप\_से]

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

क्रियान्वयन शहोउल्ड् बेलोन्ग् टो स्यन्सेट्: [कार्यान्वयन,]

गौतम शहोउल्ड् बेलोन्ग् टो स्यन्सेट्:

[गौतम\_बुद्ध,बुद्ध,गौतम,भगवान\_बुद्ध,तथागत,विश्वबोध,बुद्धदेव,सिद्धार्थ,विश्वंतर,विश्वन्तर,वीतरा  
ग,धर्मकाय,धर्मकेतु,महाश्रमण,दम,सरल,करुण]

सीआरपीएफ शहोउल्ड् बेलोन्ग् टो स्यन्सेट्: [बल,]

रोमांटिक शहोउल्ड् बेलोन्ग् टो स्यन्सेट्: [काक्टेल्,काकटेल,कॉक्टेल्,कॉकटेल]

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

झगड़ा भेलोन्स् टो स्यन्सेट्: [झगड़ा,कलह,विवाद,बखेड़ा,लड़ाई,लफड़ा,लड़ाई-

झगड़ा,झगड़ा-

लड़ाई,फसाद,फसाद,खुराफात,खुराफात,झंझट,चकल्लस,फुतूर,फुतूर,फतूर,फ़तूर,लड़ाई-

भिड़ाई,भिड़M त,भिड़न्त,टंटा,खटराग,झड़प,अड़प-

झड़प,लोचा

,रार,राड़



,षट् राग,विग्रह,चकरबा  
,अनुशय,अपडाव,अभिग्रह,अभेरा  
,रैसा,रैहर,अरवाह,अवडेर,निजा,निजा,निजाअ,निजाअ,अवरेब,ईति]

उपमुख्यमंत्री भेलोनस् टो स्यन्सेट्: [उपमुख्यमंत्री,उप\_मुख्यमंत्री,उप-  
मुख्यमंत्री,डिप्टी\_सीएम,डिप्टी\_सी\_एम]

राजस्थान भेलोनस् टो स्यन्सेट्: [राजस्थान,राजपूताना]

मेहमान शहोउल्ड् बेलोन् टो स्यन्सेट्: [कीवी,]

दोस्त शहोउल्ड् बेलोन् टो स्यन्सेट्: [मित्र,मित्र\_देव,मित्र\_देवता]

रूपया शहोउल्ड् बेलोन् टो स्यन्सेट्: [पाकिस्तानी\_रूपया,पाकिस्तानी\_रुपिया,रूपया,रुपिया]

आरती भेलोनस् टो स्यन्सेट्: [आरती,निराजन]

गंभीर शहोउल्ड् बेलोन् टो स्यन्सेट्: [नाजुक,नाजुक]

नवंबर शहोउल्ड् बेलोन् टो स्यन्सेट्: [जुलाई,]