bitext

we help AI understand humans

Synonym Data Resources

Bitext Synonym Data Resources

Bitext Synonym Data Resources are a set of synonym data resources developed to augment Wordnet®, an extremely powerful open source lexical data set of nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms (Synsets), each expressing a distinct concept. Wordnet is outdated, the last version 2.1 was released in March 2005. Bitext has developed Synonym Data Resources to update, augment and extend the value and usability of Wordnet in today's more sophisticated environment of machine learning and AI of 2020. This data has been developed to meet the highest quality standards in the field of computational linguistics and is ready to ship. Custom data solutions can also be developed on specification to support particular industry verticals, new language support and/or very specific use case requirements.

Overview

Bitext Synonym Data Resources are made of medium and high frequency entries entries that will provide value to modern day applications in the field of computational linguistics, natural language processing and machine learning - supporting use cases such as search, SEO, NLU, NLG, NLQ and conversational interfaces.

All Bitext developed entries (Synsets) have been compared with the latest release of Wordnet to confirm that they are not present in Wordnet and/or augment existing Synsets in Wordnet that are missing certain senses or synonyms.

The development of the entries have been prioritized according to their frequency in (a) Wikipedia and (b) Bitext Corpora - a tagged corpus of over 5 billion words – sourced from news, blogs, forums and social media.

In addition, lemmatization and inflection information for the synonym data set can be provided by Bitext via its proprietary Lexical Data Resources.

Specifications

- 100K+ Synonyms which are linked word meanings / senses (synsets)
- 30K+ new or enriched entries
- High confidence rate of synonym quality

Data Parameters

- Entries that are not present in WordNet
- Entries that are present in WordNet but have been enriched with new synonyms/synsets.

- Types of entries:
  - Single words (“beautiful”)
  - Multi-words (“be aware”)
  - Expressions (“hard to believe”)
  - Acronyms (“NBA”–>” National Basketball Association”)
  - Entities (“New York City”–>”NYC,Big Apple”)

- Entry-level Metadata:
  - Frequency in Wikipedia
  - Entry in Wikipedia
  - Frequency in Bitext Corpus
  - Number of linguists used for verification: 1, 2 or 3

Morphological Data

All synonyms can be linked to Bitext Lexical Data resources for lemmatization, POS and inflectional information. *more here*

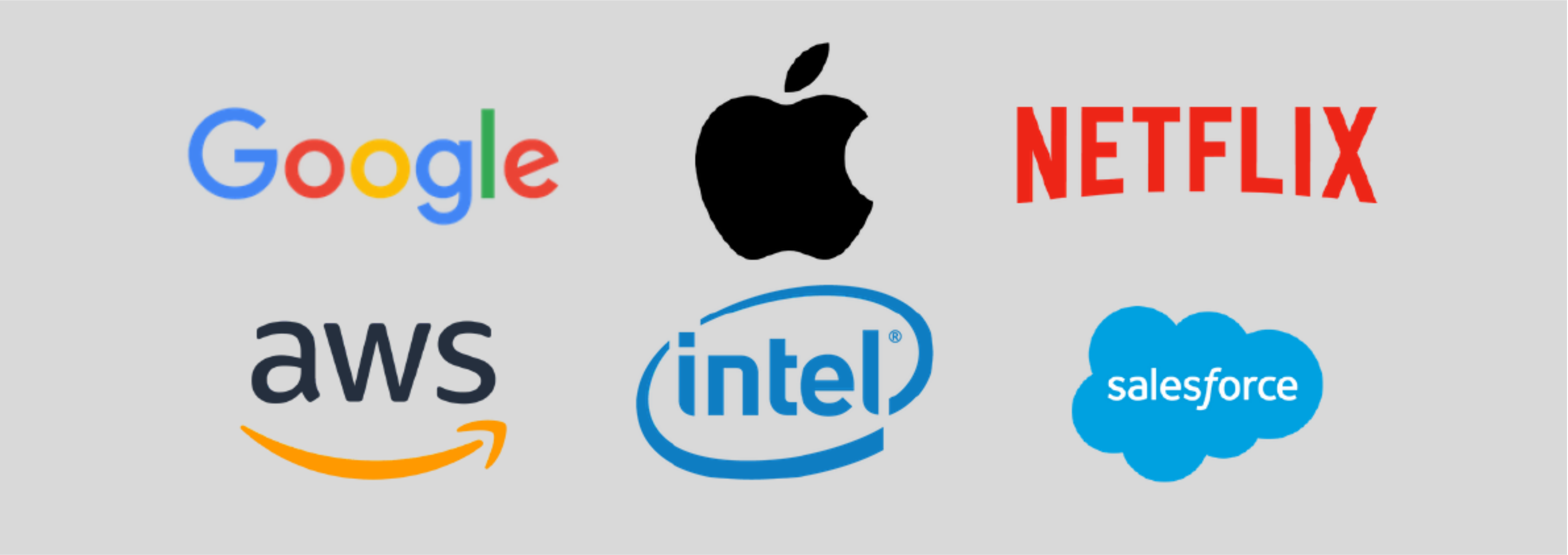
Vertical Industry Support

The current data set can be extended to include terminology for specific industries such as IT, financial services, insurance, pharmaceuticals, industrial manufacturing, energy, legal, media and entertainment, travel and hospitality, healthcare, HR, news, telecommunications, and automotive.

New Language Support

Bitext has developed Lexical Data Resources in 77 languages and 26 variants to date covering the vast majority of commercially important languages (detailed list can be found at this link). Bitext can develop new language support to augment non-English existing proprietary or open source Wordnet resources or develop synonym resources based on unique requirements provided by its customers.

Select Customers


Google
NETFLIX
aws
intel
salesforce