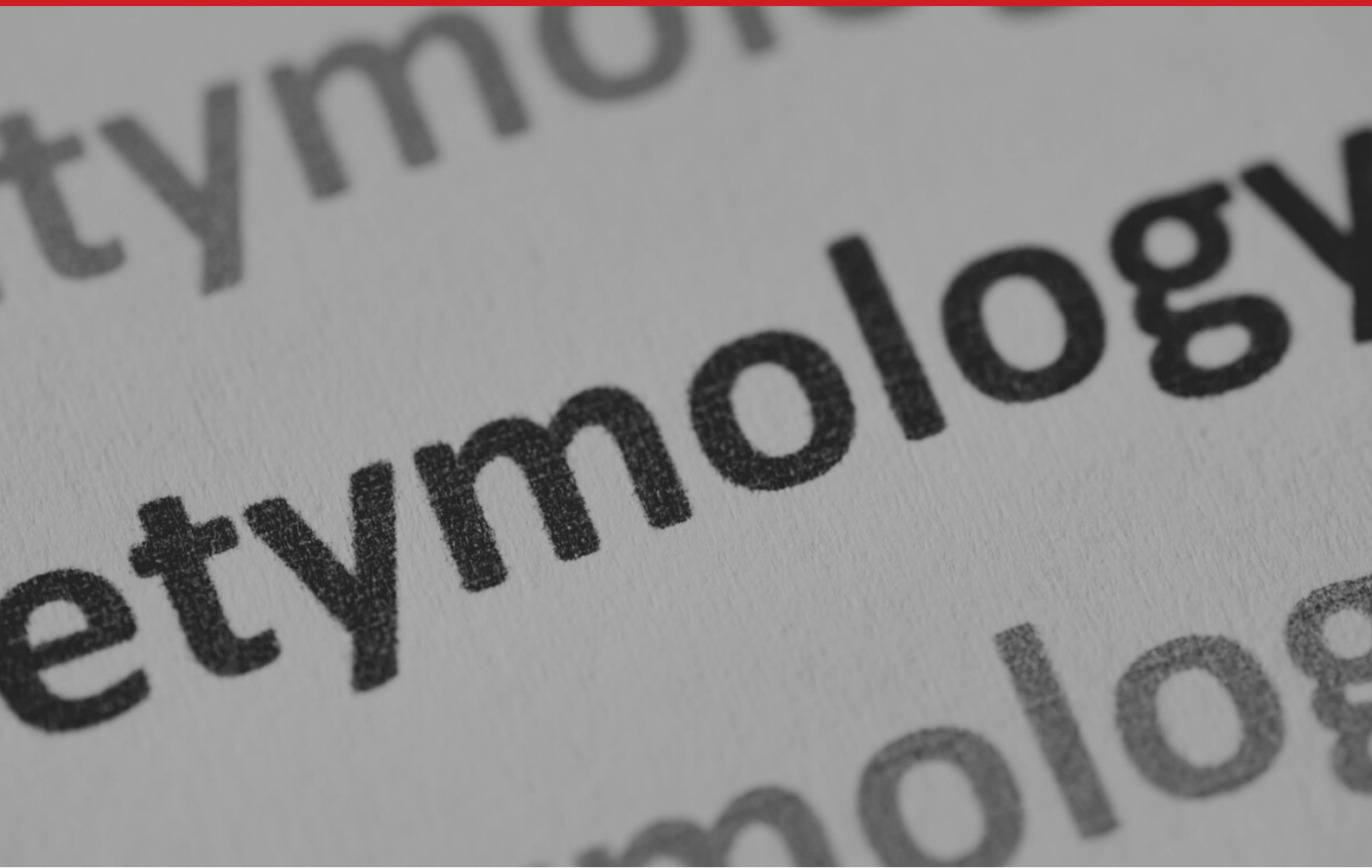
bitext we help AI understand humans

Bitext Lemmatization Benchmark



Introduction

Lemmatization is both a classic, core task of Computational Linguistics and a key component of any Natural Language Processing system. For example, it is widely used as a preprocessing step in topic modeling and it has been proven to improve the accuracy of information retrieving tasks. Today, there are numerous lemmatization tools available, for both enterprise or personal use.

This document presents a brief comparison among the NLTK stemmers and lemmatizer, the Stanford, Twinword and CST lemmatizers and the Bitext lemmatizer. The benchmark is based upon three criteria:

1. accuracy
2. performance
3. customization and maintenance

Accuracy

In terms of accuracy, stemmers pose a series of issues due to their design, which is based on general rules rather than on linguistic knowledge. Lemmatizers don't have these problems.

Also, there are big differences among lemmatizers that are key to accuracy, since they affect to very common words in every domain

Stems

One of the main issues with stemming in general, among others, is that words are stemmed to artificial words, instead of regular words. These artificial words, or stems:

>> can only be used against text stemmed with the same stemmer
>> are uninterpretable strings

Here are some examples:

Word	Stem	Lemma
easy	easi	easy
something	someth	something
Charles	charl	Charles
Husky	huski	Husky

As can be seen, lemmatization connects every word to its lemma, which is another regular word, making it a versatile and end-to-end tool.

Adjective degrees

The NLTK, Stanford, Twinword and CST lemmatizers handle cases like “easy, “namely”, “something” and proper nouns (e.g. brand names) correctly, but they do not handle derivational morphology properly.

For example, comparative and superlative forms of adjectives are not mapped to the base adjective:

Word	NLTK	Stanford	Twinword	CST	Spacy	simplemma	Bitext
easier	easy	easier	easy	easier	easy	easy	easy
cheapest	cheap	cheaper	cheap	cheaper	cheap	cheap	cheap
best	best	best	best	best	good	good	good

Pointing to the base adjective allows to make more meaningful connections.

Gerunds

Similarly, the NLTK, Stanford and Twinword lemmatizers do not map -ing verb forms to the base verb:

Word	NLTK	Stanford	Twinword	CST	Spacy	simplemma	Bitext
hammer	hammer	hammer	hammer	hammer	hammer	hammer	hammer
hammers	hammer	hammer	hammer	hammer	hammer	hammer	hammer
hammering	hammer	hammer	hammering	hammer	hammer	hammer	hammer
drill	drill	drill	drill	drill	drill	drill	drill
drills	drill	drill	drill	drill	drill	drill	drill
drilling	drill	drill	drill	drilling	drill	drill	drill

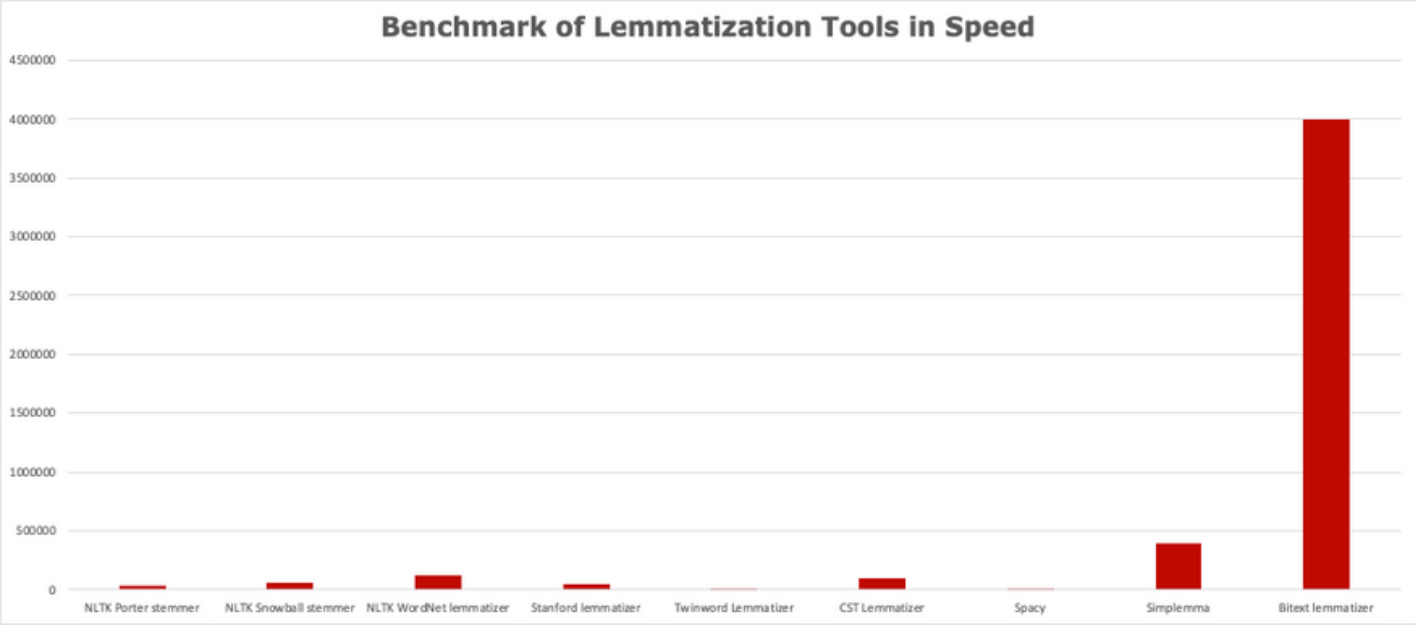
The Bitext lemmatizer handles all these cases correctly.

In short, **Bitext offers a consistent high-quality behavior across these and other phenomena.**

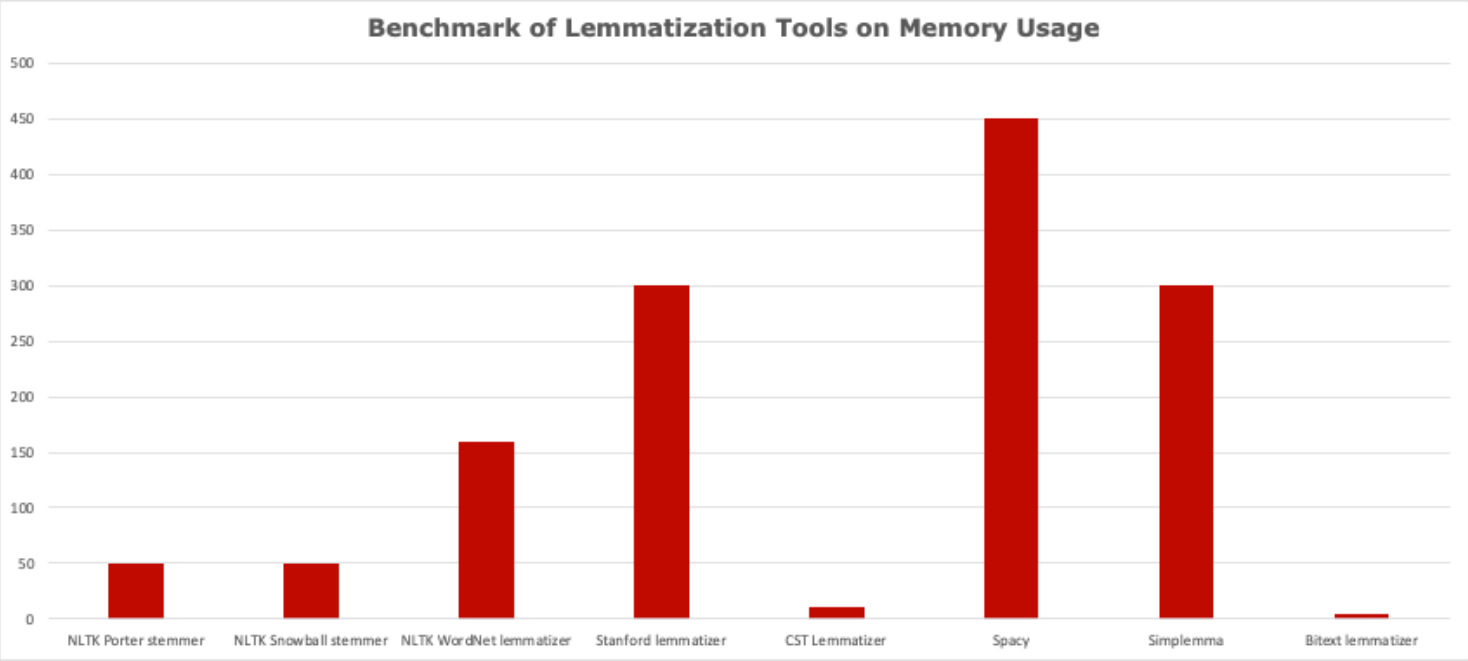
Performance

There are no standard lemmatization benchmarks or test sets, so we ran our own tests using the top 100,000 most frequent words from our base corpus.

Tool	Speed (words / sec)
NLTK Porter stemmer	40,000
NLTK Snowball stemmer	60,000
NLTK WordNet lemmatizer	120,000
Stanford lemmatizer	50,000
Twinword Lemmatizer	1,000
CST Lemmatizer	100,000
Spacy	10,000
Simplemma	400,000
Bitext Lemmatizer	4,000,000



Tool	Memory usage (MB)
NLTK Porter stemmer	50
NLTK Snowball stemmer	50
NLTK WordNet lemmatizer	160
Stanford lemmatizer	300
Twinword Lemmatizer	N/A
CST Lemmatizer	10
Spacy	450
Simplemma	300
Bitext lemmatizer	5



The Bitext lemmatizer **outperforms the rest of the lemmatizers by at least an order of magnitude.**

Customization and Maintenance

Moreover, the behavior of the Bitext lemmatizer can be easily customized by modifying or extending the dictionaries, whereas none of the other lemmatizers provide a documented procedure for modifying their behaviors.

To summarize, the Bitext lemmatizer is faster, smaller, more accurate and fully configurable.

Our Locations

Madrid, Spain
José Echegaray 8, building 3
Parque Empresarial Las Rozas
28232 Las Rozas

San Francisco, USA
541 Jefferson Ave., Ste. 100
Redwood City
CA 94063

Contact

info@bitext.com