Spell check and correction[NLP, Python]

Yash Jain · Follow
4 min read · Feb 19, 2022

In Natural Language Processing it's important that spelling errors should be as less as possible so that whatever we are making should be highly accurate. There are libraries that does this tedious task, instead of you to do all checking and correction.

We'll use **levenshtein distance, Hamming distance, Needleman-Wunsch** to check accuracy of output.

Where and why to use

> *— While having conversation with chatbots type/spell error happens and therefore context understanding becomes difficult, this is where spell correction can come handy*
>
> *— OCR post-processing — Till now no ocr gives 100% accurate results, there is always some misspell happens.*
>
> *— Fuzzy search & approximate string matching is another field where spell check/correction can be used. and there are many more applications.*

Libraries we will be using:

1. **Jamspell** `pip install jamspell` is a modern spellchecking library. It is light-weight, fast and accurate. It consider word surroundings to make better corrections. It has following features:
   It considers words surroundings (context) for better correction
   Nearly 5K words per second
   Multi-language →it's written in C++ and available for many languages with swig bindings
2. **Symspellpy** `pip install symspellpy` The Symmetric Delete spelling correction algorithm reduces the complexity of edit candidate generation and dictionary lookup for a given Damerau-Levenshtein distance. It is six orders of magnitude faster (than the standard approach with deletes + transposes + replaces + inserts) and language independent. *An average 5 letter word has about 3 million possible*

spelling errors within a maximum edit distance of 3, but SymSpell needs to generate only 25 deletes to cover them all, both at pre-calculation and at lookup time.

3. **Textblob** `pip install textblob` textblob's spelling correction is based on Peter Norvig's "How to Write a Spelling Corrector" as implemented in the pattern library.

Let's use some sample para and induce some spell errors

Error induced paragraphs:

para_1:

```
para_1 = “wherre is the love hehad dated forImuch of the past who couqdn’tread in sixthgrade and ins pired him”
```

para_2:

```
para_2 = """As far as I am abl to judg, after long attnding to the sbject, the condiions of lfe apear to act in two ways— directly on the whle organsaton or on certin parts alne and indirectly by afcting the reproducte sstem. Wit respct to te dirct action, we mst bea in mid tht in every cse, as Profesor Weismann hs latly insistd, and as I have inidently shwn in my wrk on "Variatin undr Domesticcation," thcere arae two factrs: namly, the natre of the orgnism and the natture of the condiions. The frmer sems to be much th mre importannt; foor nealy siimilar variations sometimes aris under, as far as we cn juddge, disimilar conditios; annd, on te oter hannd, disssimilar variatioons arise undder conditions which aappear to be nnearly uniiform. The efffects on tthe offspring arre ieither definnite or in definite. They maay be considdered as definnite whhen allc or neearly all thhe ofefspring off inadividuals exnposed tco ceertain conditionas duriing seveal ggenerations aree moodified
```

```
in te saame maner."""
```

para_3:

```
para_3 = """Cinderella came frm a grea family. She is the only
daughter of an affluent and widowrr duke who has rewed to provide
her witha stepmom and two stepsistrs. Cinderella’s mother died
due to illness when she was stilll a younng girl, leawing her
with a doll, faworite dress, and a pair of glasss slipppers."""
```

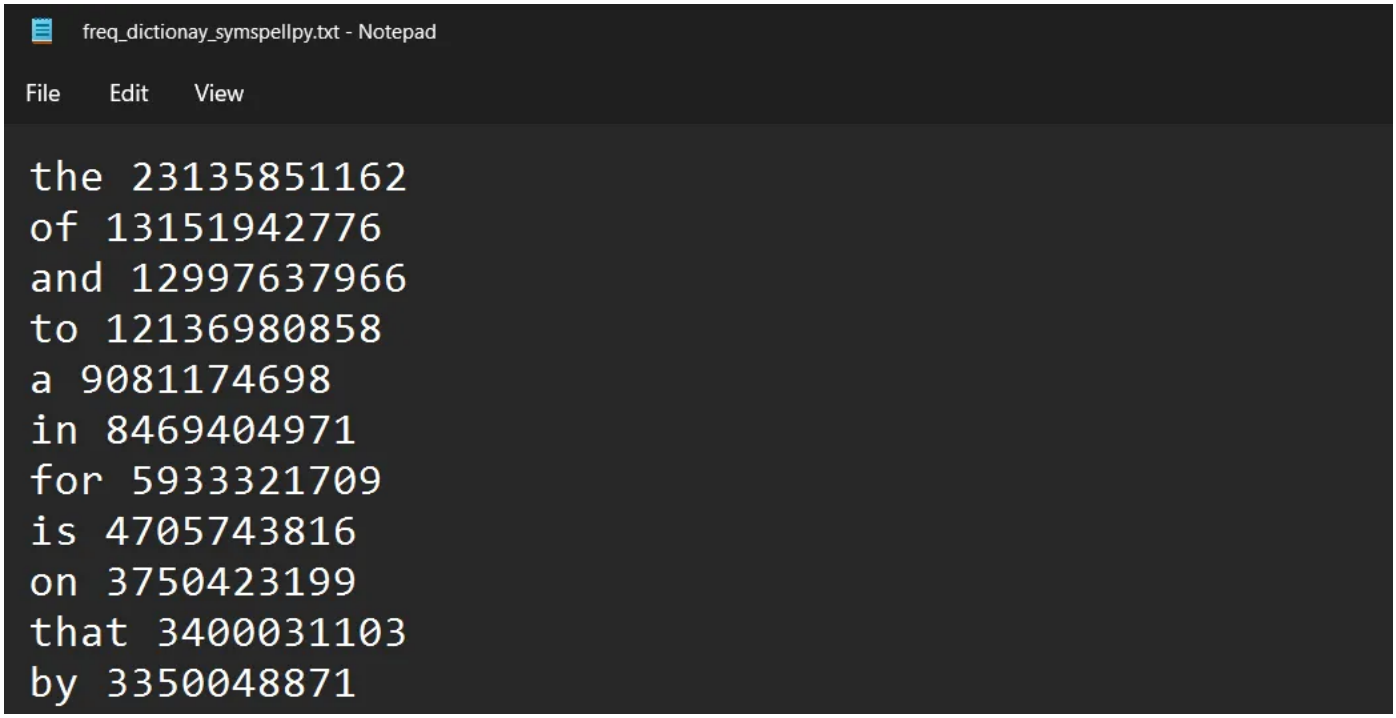
- **Jamspell**

```
!wget https://github.com/bakwc/JamSpell-
models/raw/master/en.tar.gz
!tar -xvf en.tar.gz

import jamspell
jsp = jamspell.TSpellCorrector()
assert jsp.LoadLangModel('en.bin')

jsp.FixFragment(para_1)
jsp.FixFragment(para_2)
jsp.FixFragment(para_3)
```


__Note__: You might face problem in installing and running jamspell, so i have made a docker container that exposes jamspell package as API, that you can run on your local machine. You can find instructions __here__ on how to run docker image and make a request.

- **Symspellpy**

Here I have loaded *freq_dictionay_symspellpy.txt* which is used as a corpus of words. Data is in form of 2 columns separated by space, 1st column is word, 2nd column is frequency of that word.



freq_dictionay_symspellpy.txt

You can use your own corpus. Corpus used in code can be found here

```
from symspellpy import SymSpell
symsp = SymSpell()
symsp.load_dictionary('freq_dictionay_symspellpy.txt',\
                      term_index=0, \
                      count_index=1, \
                      separator=' ')
```

Now as we have loaded our corpus of correct word in `symsp` let's try spell

correction of misspell words.

```
terms = symsp.lookup_compound(para_1,
max_edit_distance=2)
print(terms[0].term)

terms = symsp.lookup_compound(para_2,
max_edit_distance=2)
print(terms[0].term)

terms = symsp.lookup_compound(para_3,
max_edit_distance=2)
print(terms[0].term)

#max_edit_distance is the number of characters that can be
#mismatched , you can say number of wrong characters it can
tolerate
```

Output

> *Check **output** of all three para, **Correct output** and **similarity metrics** below.*

- **TextBlob**

```
from textblob import TextBlob

print(str(TextBlob(para_1).correct()))
print(str(TextBlob(para_2).correct()))
print(str(TextBlob(para_3).correct()))
```

> *Check **output** of all three para, **Correct output** and **similarity metrics** below.*

Below matrix we made: yellow are the high score **(Higher the better)**

among symspell, jamspell, textblob against para1, para2, para3 and measuring it with 3 different similarity measures with help of Damerau–Levenshtein, Hamming distance, Needleman-Wunsch. But it is just a rough estimation, you can check string output below.

Similarity check with below measures	symspell	jamspell	textblob
para1 Levenshtein	0.9505	0.9400	0.9400
para1 Hamming	0.4455	0.2400	0.2600
para1 Needleman-Wunsch	0.9505	0.9600	0.9700
para2 Levenshtein	0.9113	0.9681	0.9561
para2 Hamming	0.2004	0.1595	0.1665
para2 Needleman-Wunsch	0.9372	0.9671	0.9661
para3 Levenshtein	0.9305	0.9636	0.9636
para3 Hamming	0.1887	0.3940	0.3675
para3 Needleman-Wunsch	0.9570	0.9702	0.9768

Output of Jamspell

- Para_1 = **where is the love head dated forImuch of the past who couldn'tread in sixthgrade and ins fired him**
- Para_2 = **As far as I am able to judg, after long attending to the subject, the conditions of life appear to act in two ways-directly on the whole organisation or on certain parts alone and indirectly by acting the reproduce system. Wit respect to the direct action, we mst bea in mid the in every case, as Professor Weismann hs lately insisted, and as I have incident shown in my work on "Variation under Domestication," there are two factors: namely, the nature of the organism and the nature of the conditions. The former seems to be much th mre important; for nearly similar variations sometimes aris under, as far as we can judge, dissimilar conditions; and, on the other hand, dissimilar**

variations arise under conditions which appear to be nearly uniform. The effects on the offspring arre either definite or in definite. They may be considered as definite when all or nearly all the offspring off individuals exposed to certain conditions during several generations are modified in the same manner.

- Para_3 = **Cinderella came from a great family. She iss te only daughter of an affluent and widower duke who has reed to provide heer with stepmum and two stepsistrs. Cinderella's mother died due to illness when she was still a young girl, leading her with a doll, favorite dress, and a pair of glasses slippers.**

Output of Symspell

- Para_1 = **where is they love head dated for much of they past who couldn't read in sixth grade and inspired him**
- Para_2 = **as far as i am abl to judge after long attending to they subject they conditions of life appear to act in two ways directly on they while organs ton or on certain parts alone and indirectly by acting they reproduce system wit respect to to direct action we most be in mid that in every use as professor weizmann is lately insist and as i have evidently shown in my work on variation under domestication there are two factors namely theatre of they organism and they nature of they conditions they former seems to be much there important for nearly similar variations sometimes arts under as far as we in judge dissimilar conditions and on to other hand dissimilar variations arise under conditions which appear to be nearly uniform they effects on tithe offspring are either definite or in definite they may be considered as definite when all or nearly all thee offspring off individuals exposed to certain conditions during several generations are modified in to same manner**
- Para_3 = **cinderella came from a great family she issue only daughter of an affluent and widower duke who has reed to provide here with stepmom and two stepsisters cinderellas mother died due to illness when she was still a young girl leading or with a dolls favourite dress and a pair of glass slippers**

Output of Textblob

- Para_1 = **where is the love head dated forImuch of the past who couldn'tread in sixthgrade and in tired him**
- Para_2 = **Is far as I am all to judge, after long attending to the subject, the conditions of life appear to act in two ways-directly on the while organisation or on certain parts alone and indirectly by acting the reproduce system. It**

respect to te direct action, we must be in mid the in every case, as Professor Weismann he lately insisted, and as I have evidently shown in my work on "Variation under Domesticcation," there are two facts: namely, the nature of the organism and the nature of the conditions. The former seems to be much th are important; for nearly similar variations sometimes arms under, as far as we in judge, similar condition; and, on te other hand, disssimilar variations arise under conditions which appear to be nearly uniform. The effects on the offspring are either definite or in definite. They may be considered as definite when all or nearly all the offspring off individuals exposed to certain conditions during several generations are modified in te same manner.

- Para_3 = **Cinderella came from a great family. The iss te only daughter of an affluent and widower duke who has reed to provide her with sternum and two stepsistrs. Cinderella's mother died due to illness when she was still a young girl, leaving he with a doll, favorite dress, and a pair of glass slippers.**

Correct Strings

Correctpara1 = **where is the love he had dated for much of the past who couldn't read in sixthgrade and inspired him**

Correctpara2 = **As far as I am able to judge, after long attending to the subject, the condiions of life appear to act in two ways-directly on the whole organisaton or on certain parts alone and indirectly by affcting the reproductive system. With respect to te direct action, we must beat in mid that in every case, as Professor Weismann has lately insisted, and as I have inidently shown in my work on "Variation undr Domestication," there are two factors: namely, the nature of the organism and the nature of the conditions. The former seems to be much more importannt; for nearly siimilar variations sometimes aris under, as far as we can judge, disimilar conditions; and, on the other hand, dissimilar variations arise under conditions which appear to be nearly uniform. The efffects on the offspring are either definite or indefinite. They may be considered as definite when all or nearly all the offspring off individuals exposed to ceertain conditions during**

We have looked at three different spell correction libraries, every output has some erroneous text remains. A metrics for comparison. There is little difference in every package output. You might have to experiment and go through algorithm behind the packages to pick library that suits

your need.

Spell Check Python NLP Spelling Spell Correction

Written by Yash Jain

117 Followers · 20 Following

Follow

Data Scientist/ Data Engineer at IBM | Alumnus of @niituniversity | Natural Language Processing | Pronouns: He, Him, His

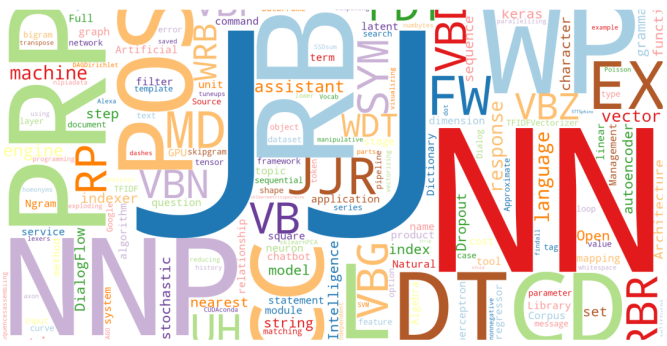
More from Yash Jain

Yash Jain

Stopwords [NLP, Python]

Stop words are common words in any language that occur with a high frequency...

Feb 23, 2022 9

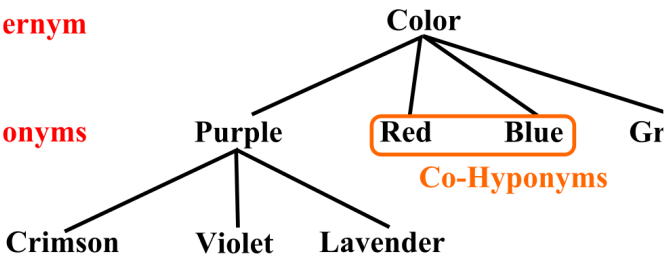


Yash Jain

POS Tagging [NLP, Python]

POS tagging is important to get an idea that which parts of speech does tokens...

Feb 27, 2022 19 1

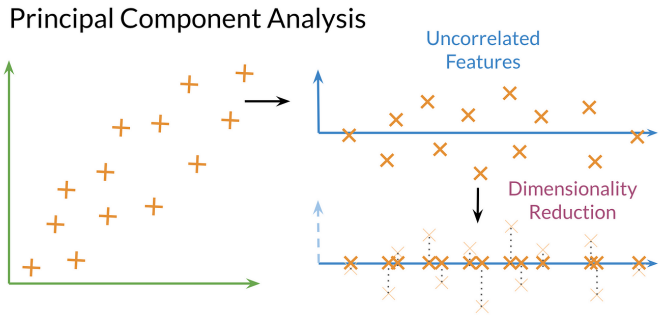


Yash Jain

Understanding Hypernyms and Hyponyms in NLP using Python...

Hypernyms and hyponyms are two important concepts in natural language...

Mar 22, 2023 6



Yash Jain

Principal Component Analysis (PCA) [NLP, Python]

It's a common practice of reducing the dimension, PCA is an unsupervised...

May 9, 2022 8

See all from Yash Jain

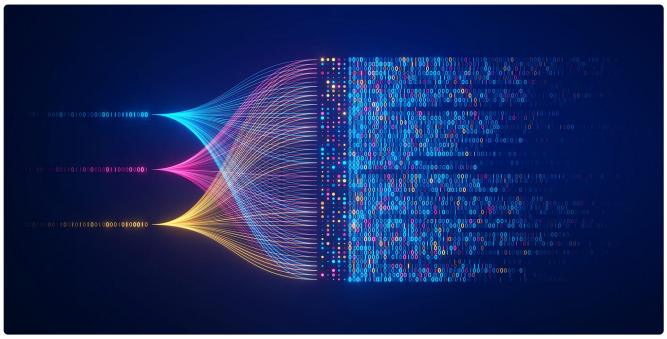
Recommended from Medium

In Towards Data Science by Youness Mansar

Building a Reliable Text Classification Pipeline with LLM...

Overcoming common challenges in LLM-based text classification

5d ago 422 4



In inganalytics.com/inganalytics by Max Baak

Yet even faster string comparison!

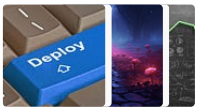
Much improved version of sparse_dot_topn now released, check out v1.1.1

May 26

Lists

Coding & Development

11 stories · 906 saves



Predictive Modeling w/ Python

20 stories · 1677 saves

Practical Guides to Machine Learning

10 stories · 2029 saves

Natural Language Processing

1814 stories · 1425 saves

In Stackademic by Abdur Rahman

Python is No More The King of Data Science

5 Reasons Why Python is Losing Its Crown

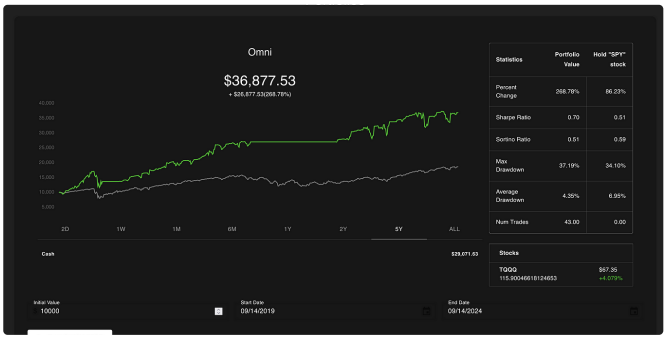
Oct 23 6.5K 28

In Python's Gurus by Ime Inyang Jnr.

Performing Resumé Analysis using NER with Cosine Similarity

Expediting the Hiring Process using Named Entity Recognition (NER) Systems

Jun 24 56 1

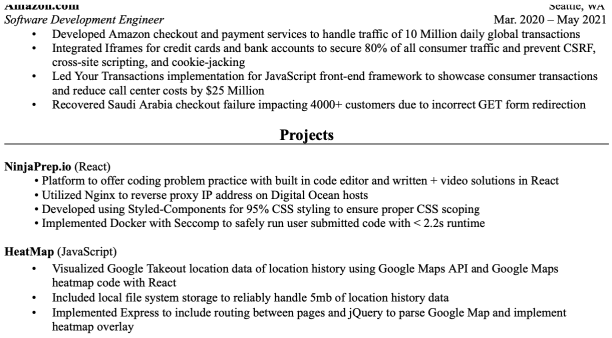


In DataDrivenInvestor by Austin Starks

I used OpenAI's o1 model to develop a trading strategy. It is...

It literally took one try. I was shocked.

Sep 15 6.2K 154



Software Development Engineer Mar. 2020 – May 2021

- Developed Amazon checkout and payment services to handle traffic of 10 Million daily global transactions
- Integrated Iframes for credit cards and bank accounts to secure 80% of all consumer traffic and prevent CSRF, cross-site scripting, and cookie-jacking
- Led Your Transactions implementation for JavaScript front-end framework to showcase consumer transactions and reduce call center costs by $25 Million
- Recovered Saudi Arabia checkout failure impacting 4000+ customers due to incorrect GET form redirection

Projects

NinjaPrep.io (React)
- Platform to offer coding problem practice with built in code editor and written + video solutions in React
- Utilized Nginx to reverse proxy IP address on Digital Ocean hosts
- Developed using Styled-Components for 95% CSS styling to ensure proper CSS scoping
- Implemented Docker with Seccomp to safely run user submitted code with < 2.2s runtime

HeatMap (JavaScript)
- Visualized Google Takeout location data of location history using Google Maps API and Google Maps heatmap code with React
- Included local file system storage to reliably handle 5mb of location history data
- Implemented Express to include routing between pages and jQuery to parse Google Map and implement heatmap overlay

In Level Up Coding by Alexander Nguyen

The resume that got a software engineer a $300,000 job at...

1-page. Well-formatted.

Jun 1 25K 510

See more recommendations