Often when dealing with long sequences of text you’ll want to break those sequences up and extract individual keywords to perform a search, or query a database.

If the input text is natural language you most likely don’t want to query your database with every single word — instead, you probably want to choose a set of unique keywords from your input and perform an efficient search using those words or word phrases.

This task is known as keyword extraction and thanks to production grade NLP tools like [Spacy](https://spacy.io/) it can be achieved in just a couple of lines of Python. In this article we will cover:

* how to build a simple and robust keyword extraction tool using Spacy
* how to handle spelling mistakes and find fuzzy matches for a given keyword (token) using [fuzzyWuzzy](https://github.com/seatgeek/fuzzywuzzy)
* how to wrap both of these functions up into REST API endpoints with [Flask](https://flask.palletsprojects.com/en/1.1.x/)

This lightweight API is intended to be a general purpose keyword service for a number of use cases. You can of course also build any of Spacy’s numerous NLP functions into this API using the same general structure.

Before we start, make sure to run: pip install flask flask-cors spacy fuzzywuzzy to install all the required packages.

Keyword Extraction with Spacy

For the keyword extraction function, we will use two of Spacy’s central ideas— the core language model and document object.

Spacy [Core language models](https://spacy.io/models) are:

General-purpose pretrained models to predict named entities, part-of-speech tags and syntactic dependencies. Can be used out-of-the-box and fine-tuned on more specific data.¹

The Spacy [document](https://spacy.io/api/doc) object is:

A container for accessing linguistic annotations…(and) is an array of token structs²

So with the creation of a document object created via the model we are given access to a number of very useful (and powerful) NLP derived attributes and functions including part-of-speech tags and noun chunks which will be central to the functionality of the keyword extractor.

Download Language Model

With Spacy we must first download the language model we would like to use. As of today Spacy’s current version 2.2.4 has language models for 10 different languages, all in varying sizes. I will be using the small version of the English Core model. I chose the small model as I had issues with the size of the large model in memory for Heroku deployment. Depending on where/how you deploy this model you may be able to use the large model. To download the language model using Spacy’s CLI run the following command in your terminal: