# Design

**Tokenization & morphology analysis**

* Sentence ending flag

. ? ! \n

**Syntactic parse**

* **Spacy**
* **Morphology analysis**

**Semantic parse**

* **NER (spacy)**

**Question parse**

* when, who, where, whom, why -> parse directly
* what, which, how -> question focus (the first noun group that is not the word ”name”)

**sentence representation**

* sentence2vec

# Milestone

**Milestone 1**

* overlap(question, sentence\_i)
* use pre-trained embedding, sim(question, sentence\_i) based on sentence2vec

**Milestone 2**

* root(question) in (sub)root(sentence)
* parse questions

**Milestone 3**

* coreference
  + multi-pass system (very competitive and easy to use)
* extract sentence span

# Data format

**Story\_data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Story\_id** | **headline** | **date** | **Story (type of doc)** |
| 1 |  | 1 | 1 |
| 1 |  | 1 | 2 |
| 1 |  | 2 | 1 |
| 2 |  |  |  |

**Question\_and\_ans\_data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Story\_id** | **Question\_id** | **question** | **Difficulty** | **focus** | **type** | **answer** | **Answer\_pred** |
|  | 1 | Which city is located in …. |  | city | LOCATION |  |  |
|  |  | When |  |  | DATE |  |  |
|  |  | Who is Yu Zhu? |  |  |  |  |  |
|  |  | Who killed her? |  |  |  |  |  |

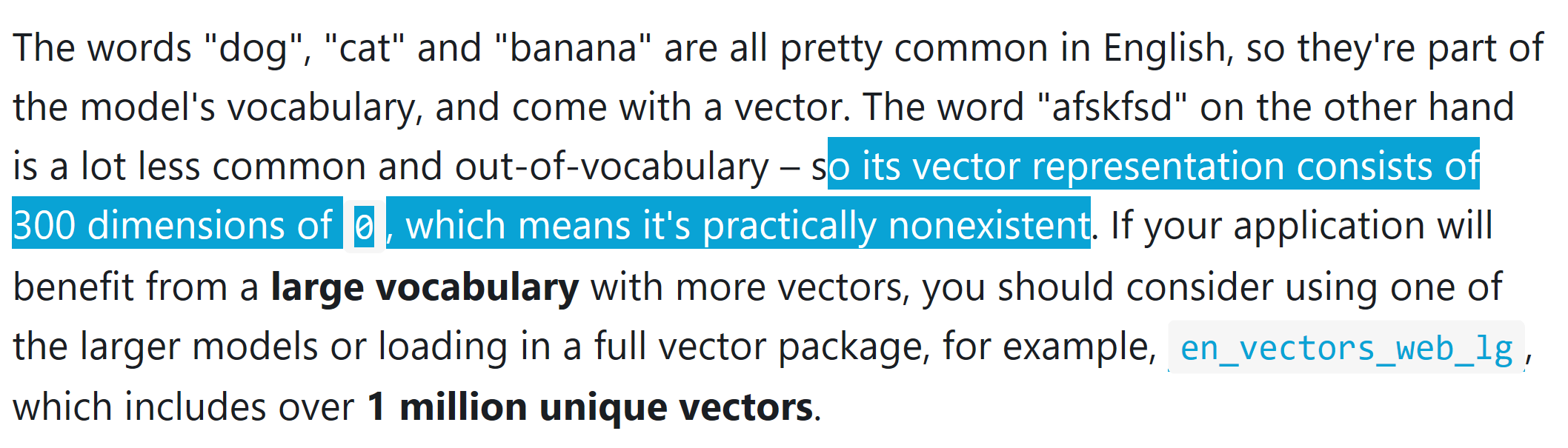
# Resources

**Word2vec**

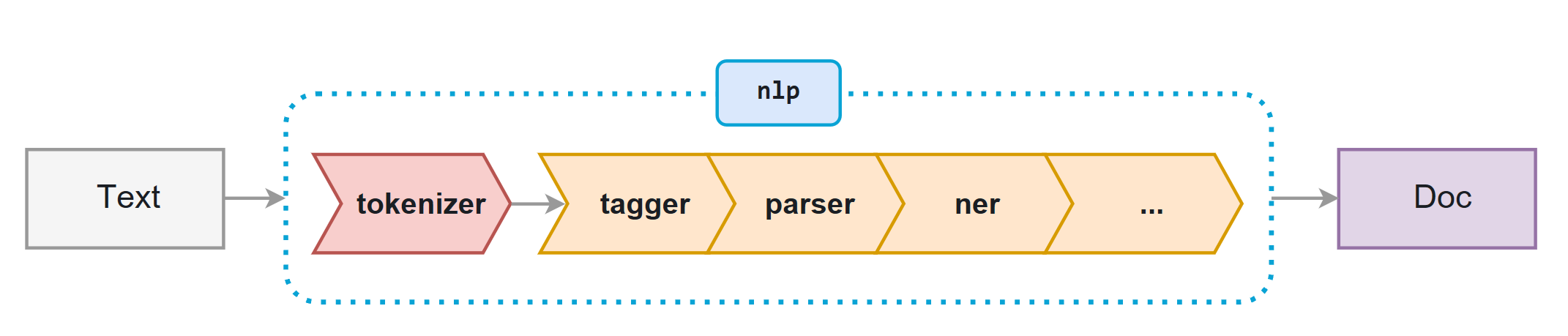
* [spaCy word vector](https://spacy.io/models/en#section-en_vectors_web_lg) (1.1m keys, 1.1m unique vectors, 300 dimensions)

# Notes

* BSD. Boundary of sentence detection.
* Unseen words

[](https://spacy.io/usage/spacy-101#vectors-similarity)

* Pipeline



* WordSim(work, works) = 0.83! We need to compare the lemmas!!!