Search Engine

A simple search engine with ranking and graghic interface

Written 100% in python

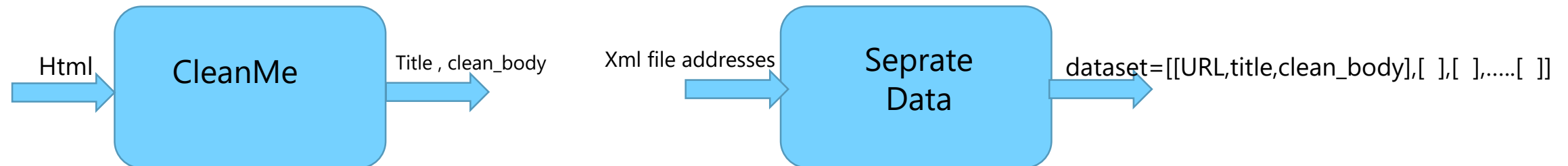# Parsing

Reading .xml files

❖ **liberary** :ElementTree

Separate title, body and url

❖ liberary : BeautifulSoup

❖ Challenge: some websites didn't have title

Remove tags and sort text

❖ liberary : BeautifulSoup

❖ Challenge: Some signs that were not tagged remain

Functions:

Html → [CleanMe] → Title , clean_body

Xml file addresses → [Seprate Data] → dataset=[[URL,title,clean_body],[ ],[ ],…..[ ]]

Indexing

Calculating TF-IDF :

❖ Title : title_tfidf [(doc1,token : tfidf),( ), ...()]

❖ Body : body_tfidf [(doc1,token : tfidf),( ), ...()]

Formula :

TF-IDF = body_tf-idf * body_weight + title_tf-idf*title_weight

Alpha : weight of body

1- Alpha =weight of title

Ranking

Ranking:

❖ Matching Score
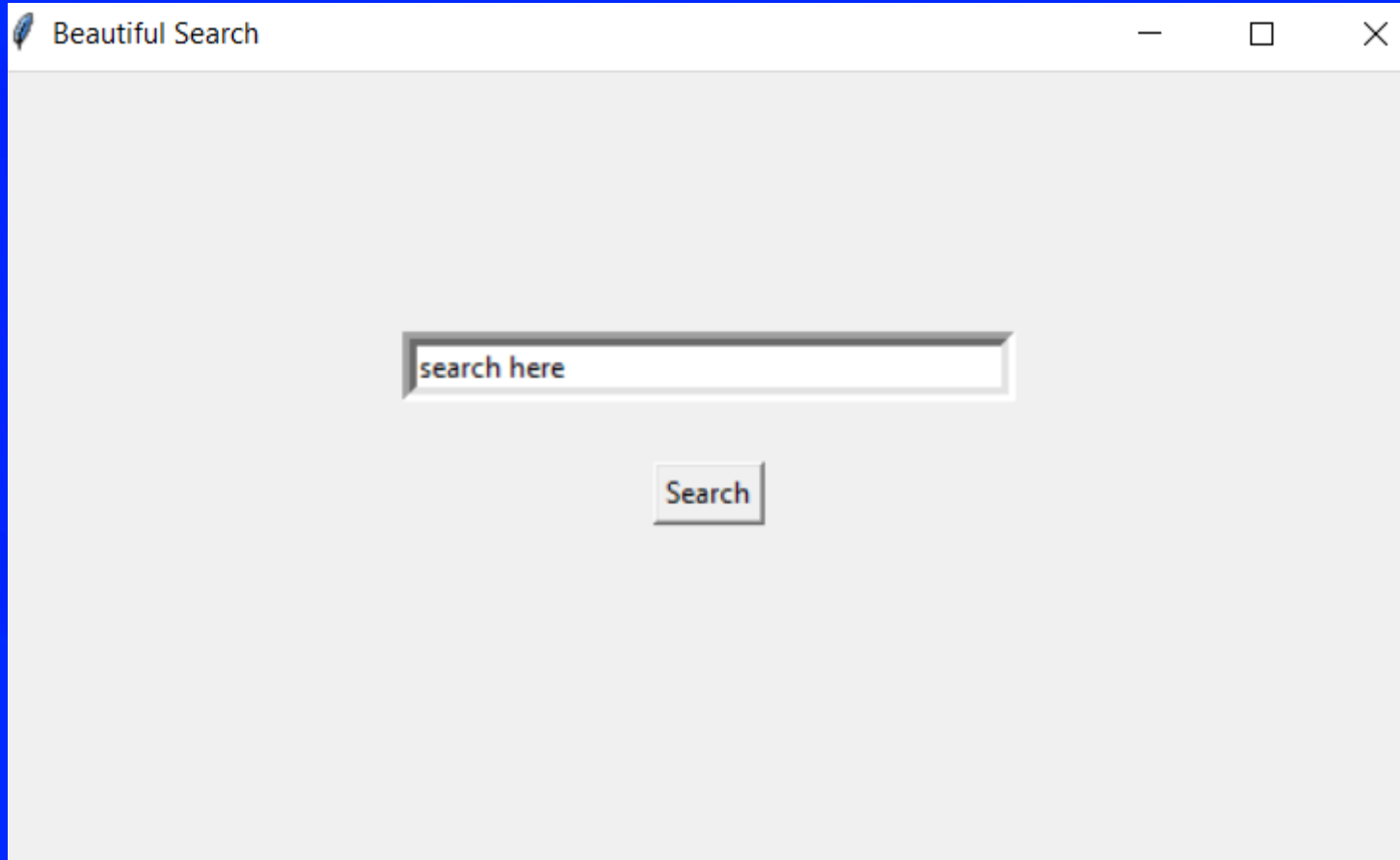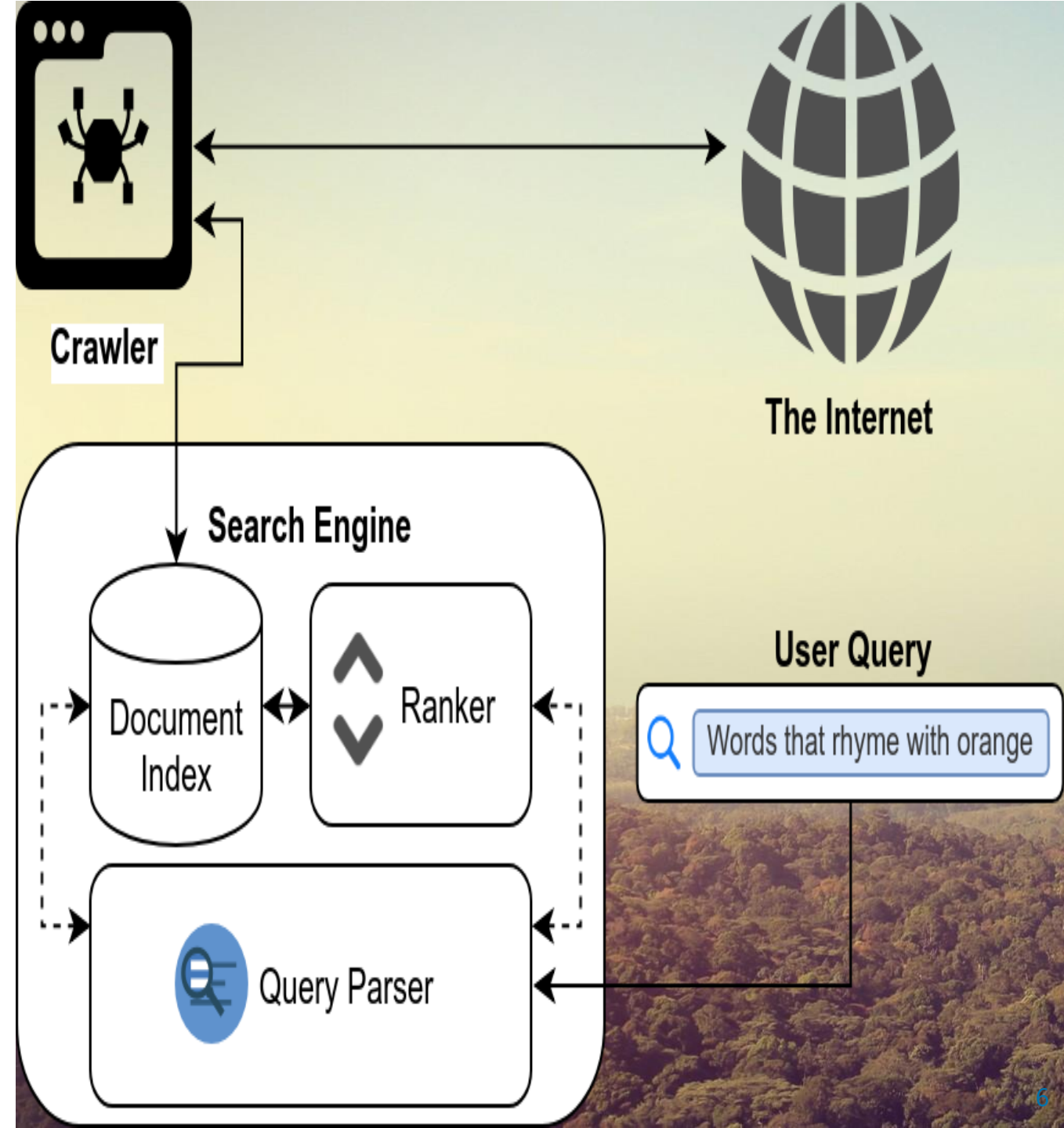
✓ Cosine Similarity

GUI
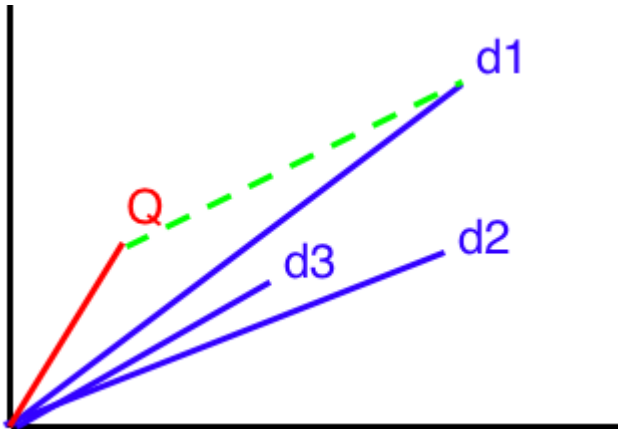
1. **The sites are Persian**

2. **The content of the sites is the same**

3. **Most of the links are dead**

4. **Data volume**

Crawler

The Internet

Search Engine

Document Index

Ranker

Query Parser

User Query

Words that rhyme with orange

Thank you

Developer: Farnaz Farhand