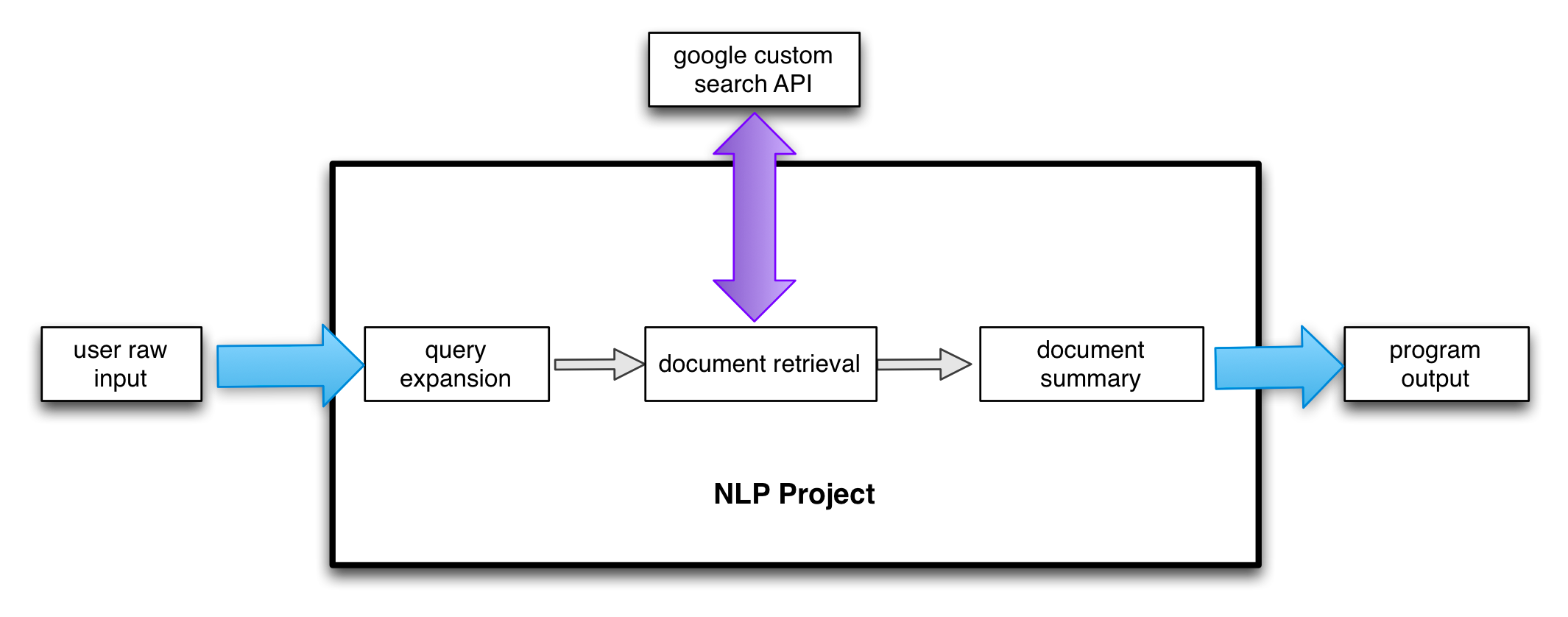
NLP Project Documentation

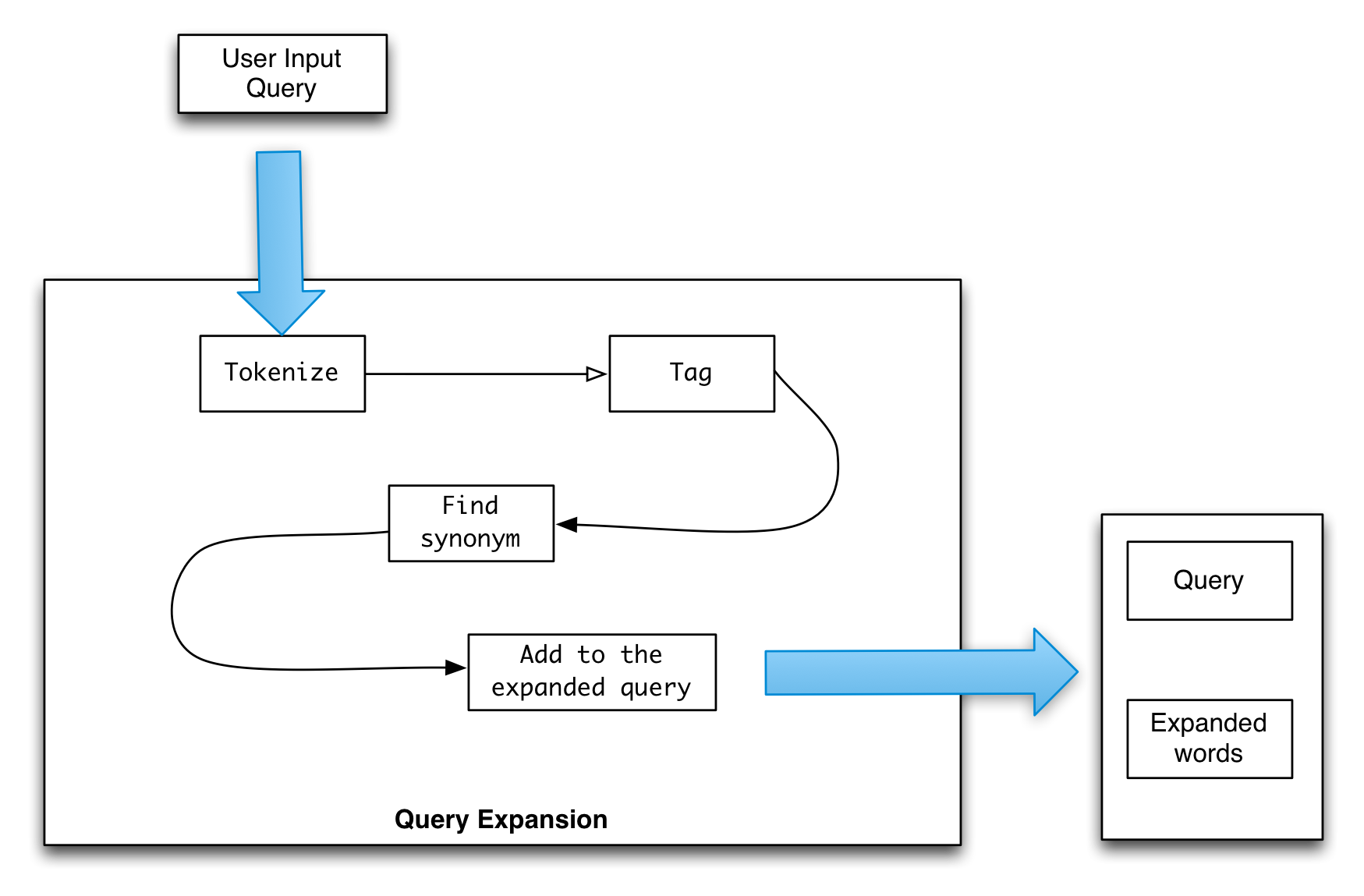
**The Project requirement**

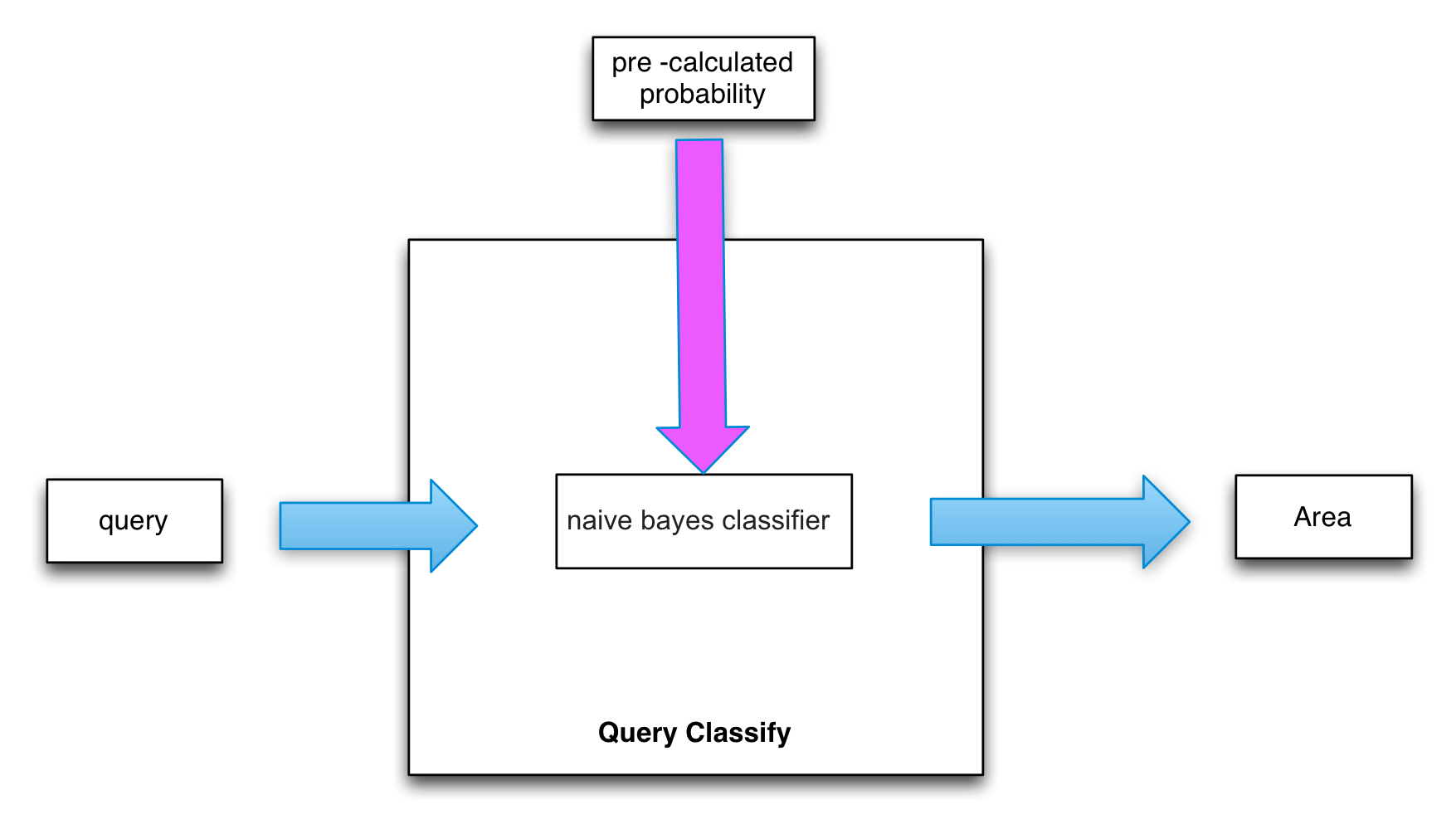
Web application: a web page for submitting search queries and display results. This page should be hosted on a server, either local or on the web. It should have the functionality of submitting the search query to the service for getting search results and annotating with snippets and getting results from this service. The web application can be implemented in any language. Snippet generating service will consist of three parts: document retrieval service, document analysis service, and snippet delivery service. The document retrieval service will accept search query requests, make calls to Google or Bing API, retrieve line from the API, and retrieve the documents that these line point to. The document analysis service will analyze their documents to generate snippets.



**System analysis**

* User input
  + The raw data provided by user
  + Class name: UserInput
* Query expansion
  + Expanded query use nltk wordnet interface to find the synonym for each non STOP word in query
  + Classify the query to one of the three area using naïve Bayes
  + Class name: QueryClassifier , QueryExpansionProcess





* Document retrieval
  + Passed the query to the Goolge API and retrieve the text data by using BeautifulSoup
  + Class name: Search, Clean
* Document summary
  + We used the tf-idf to summary the retrieved document (the details are showing table 1)
  + Class name: Summary

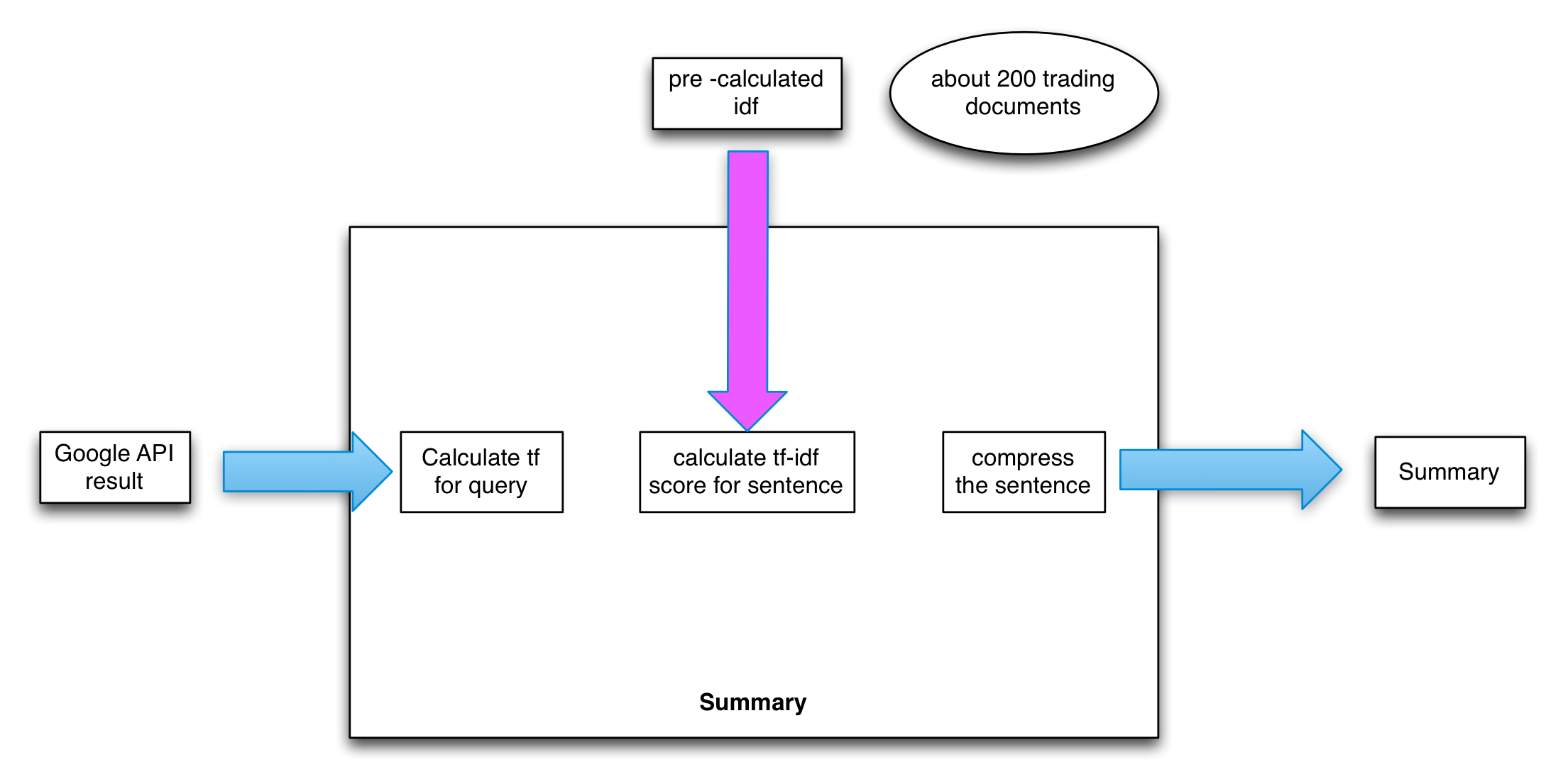


Table 1 : The algorithm and steps for summary:

|  |
| --- |
| 1. tf-idf |
| 1. idf = log(total documents / 1+ number of document contain the word) |
| 1. The documents for training idf got from different websites and storage in training folder |
| 1. tf = 0.5 + (0.5 \*word frequency/ document words count) |
| 1. tf-idf socre = tf \* idf |
| 1. If the word is contained in the query or expanded query the weight will be increase by 5% |
| 1. Sentence weight = sum of tf-idf score of all the word in the sentence / number of words in the sentence |
| 1. The return summary is the two have the highest weight sentence |