**CS5560 Knowledge Discovery Management**

**LAB ASSIGNMENT #3**

**Name: Lava Kumar Surparaju**

**Class ID: 27**

**Lab 3 Assignment:**

1. ***In Class Question***

**Generate the output (changes or transformations in the data) manually when the following tasks are applied on the input text. Show your output in details.**

***Input***:

*Doc#1: The dog saw John in the park*

*Doc#2: The little bear saw the fine fat trout in the rocky brook.*

*Doc#3: The dog started chasing John.*

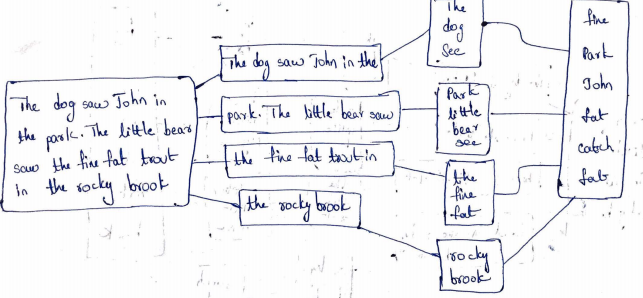
*Doc#4: The little bear caught a fish in the rocky brook.*

***Tasks:***

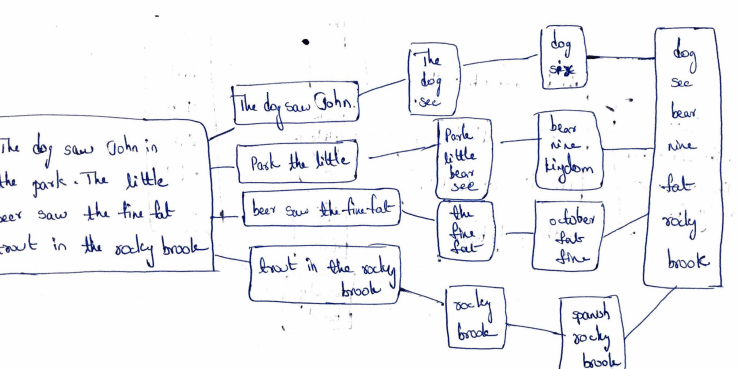
* 1. ***Find out the top TF-IDF words for the above input.***
     + *Fat*
     + *Finr*
     + *Park*
     + *Caught*
     + *Chasing*
     + *Started*
     + *Trout*
     + *Fish*
     + *Saw*
     + *Rocky*
     + *Little*
     + *Brook*
     + *Bear*
     + *Dog*
     + *John*
  2. ***Find out the top TF-IDF words for the lemmatized input***
     + *fat*
     + *fine*
     + *park*
     + *catch*
     + *chase*
     + *start*
     + *trout*
     + *fish*
     + *see*
     + *rocky*
     + *little*
     + *brook*
     + *bear*
     + *dog*
     + *John*
  3. ***Find out the top TF-IDF words for the n-gram based input.***
     + *chase*
     + *fat*
     + *park*
     + *a*
     + *chase*
     + *Fine*
     + *fish*
     + *Catch*
     + *start*
     + *fine*
     + *park*
     + *fish*
     + *trout*
     + *fat*
     + *brook*
     + *trout*
     + *catch*
     + *start*
     + *rocky*
     + *John*

***Draw Map-Reduce Diagram for the TF-IDF and W2V (similar to Lab2)***

***TF-IDF:***



***W2V:***

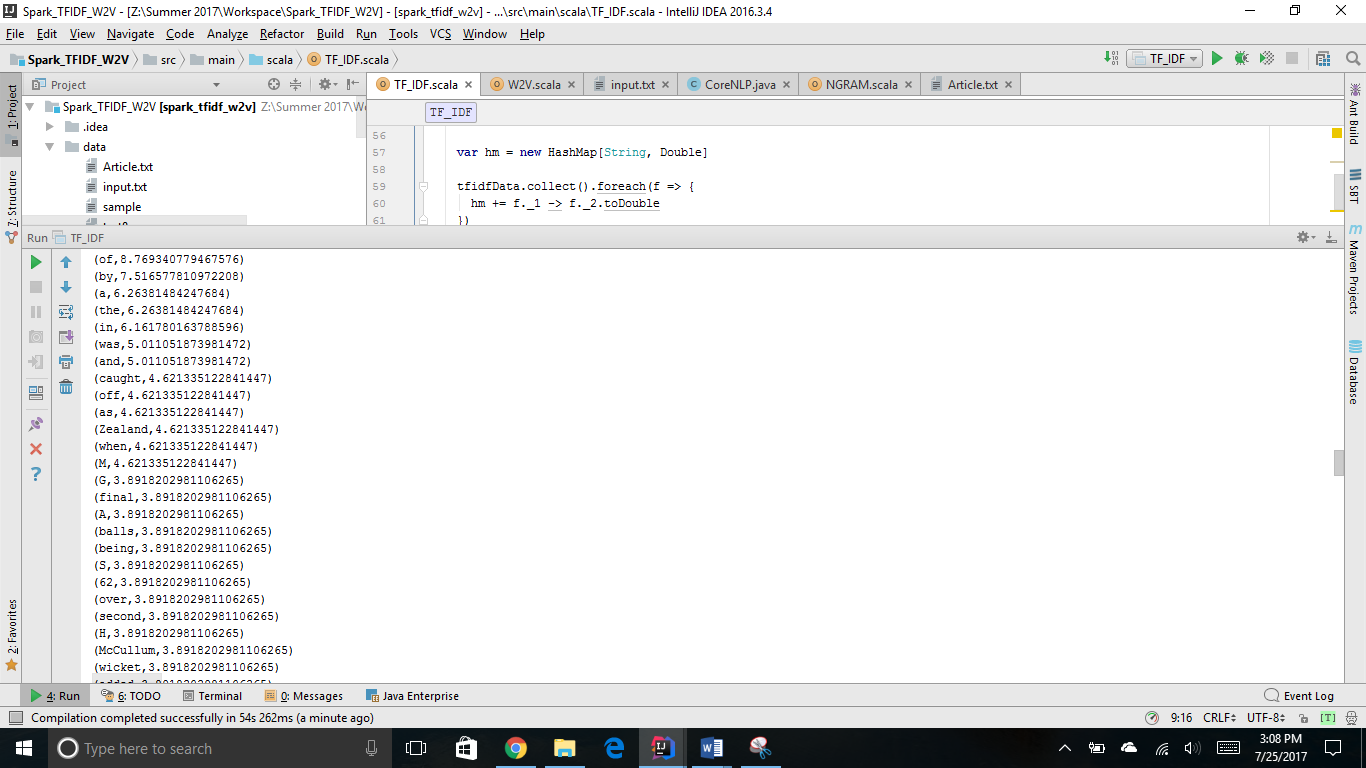


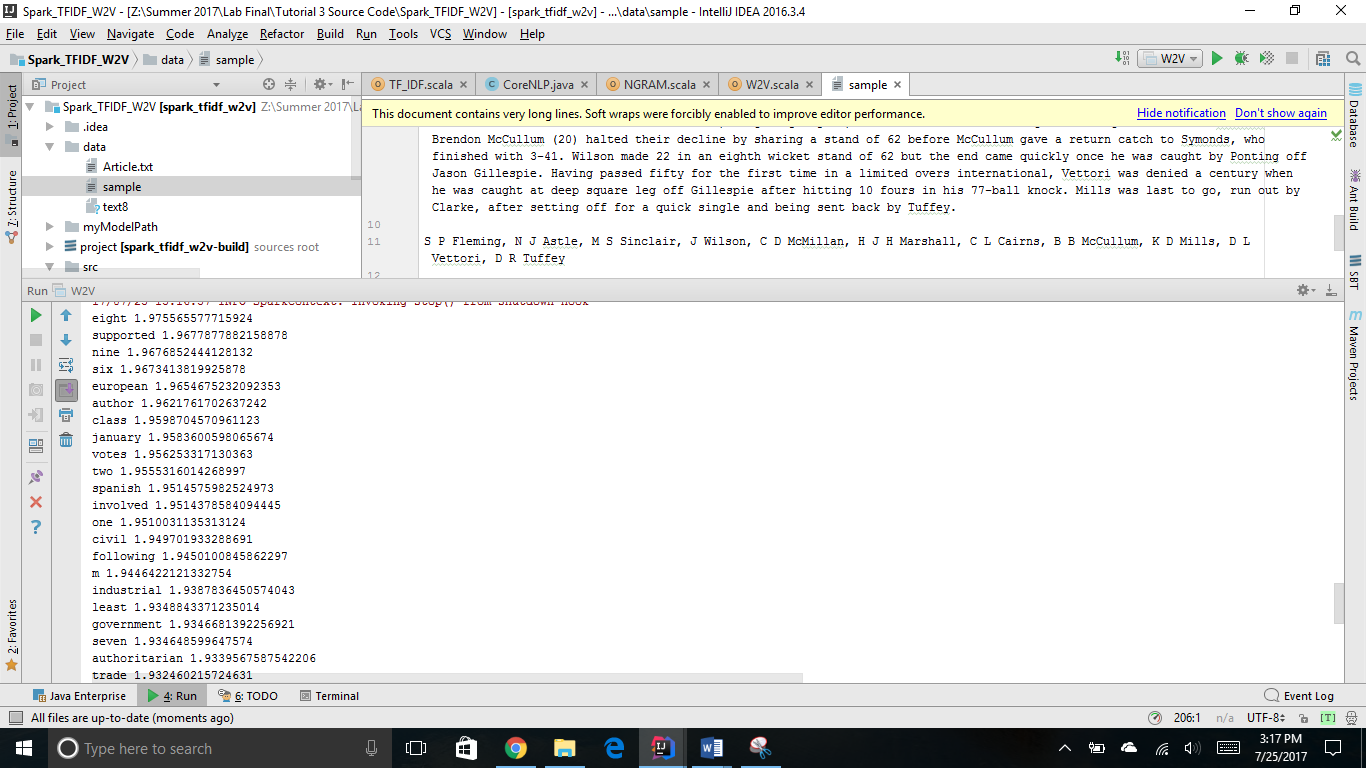
1. ***In Class Question***

**Write a simple spark program to read a dataset and find the W2V Synonyms for the Top TF-IDF Words**

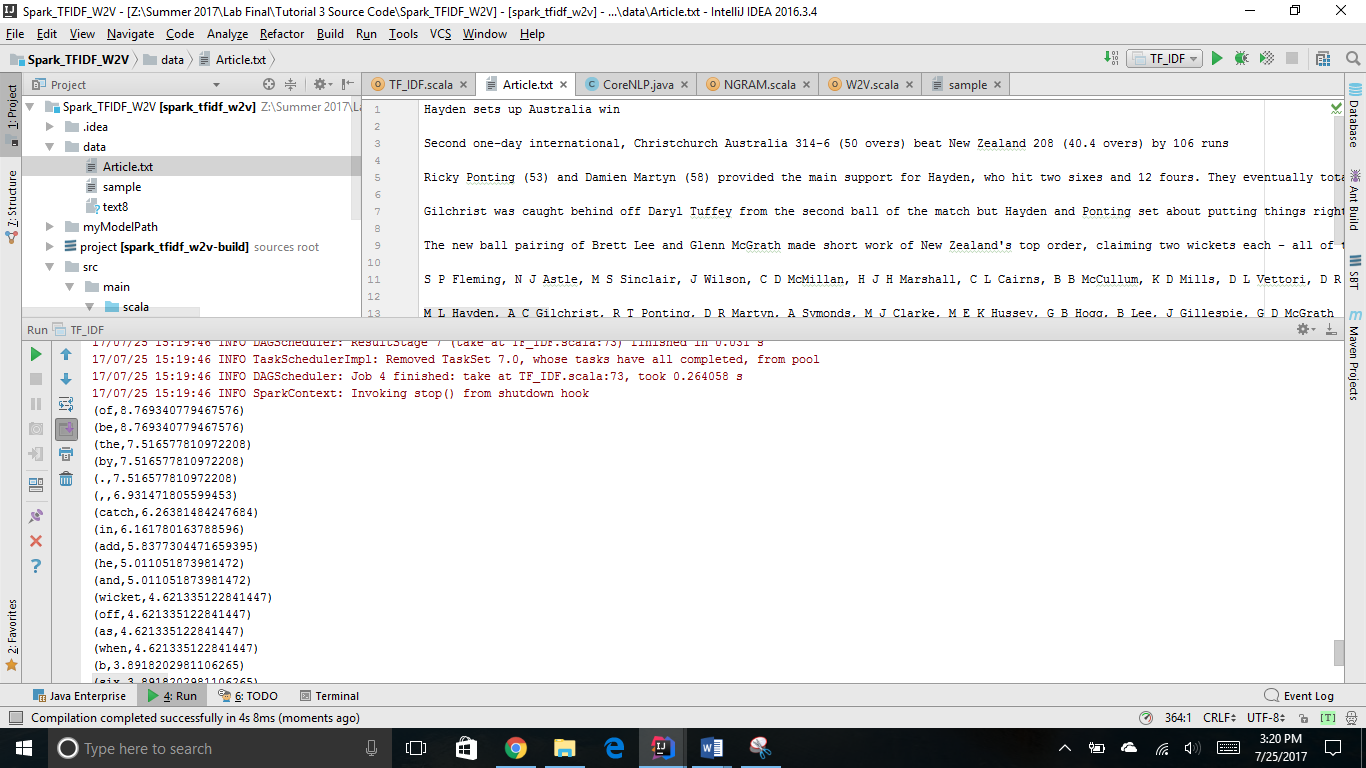
* 1. Try without NLP

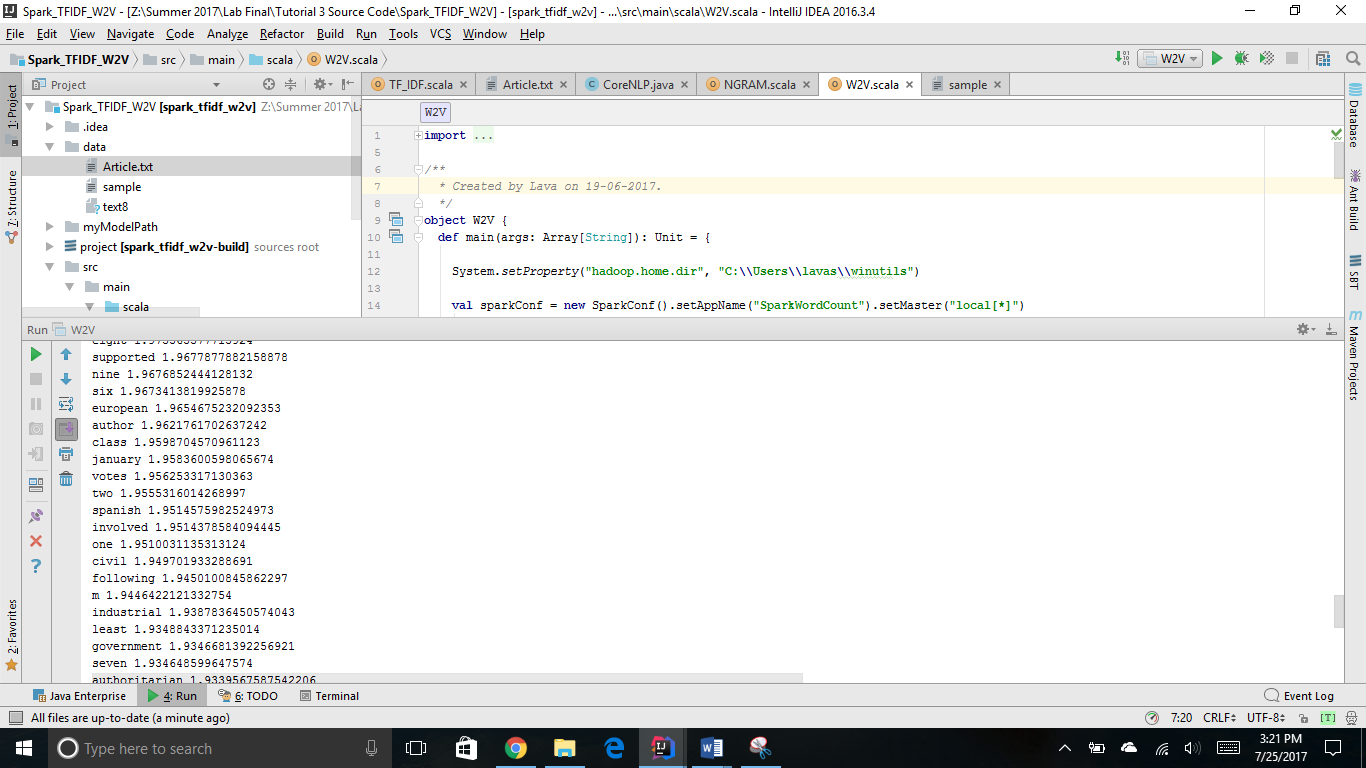




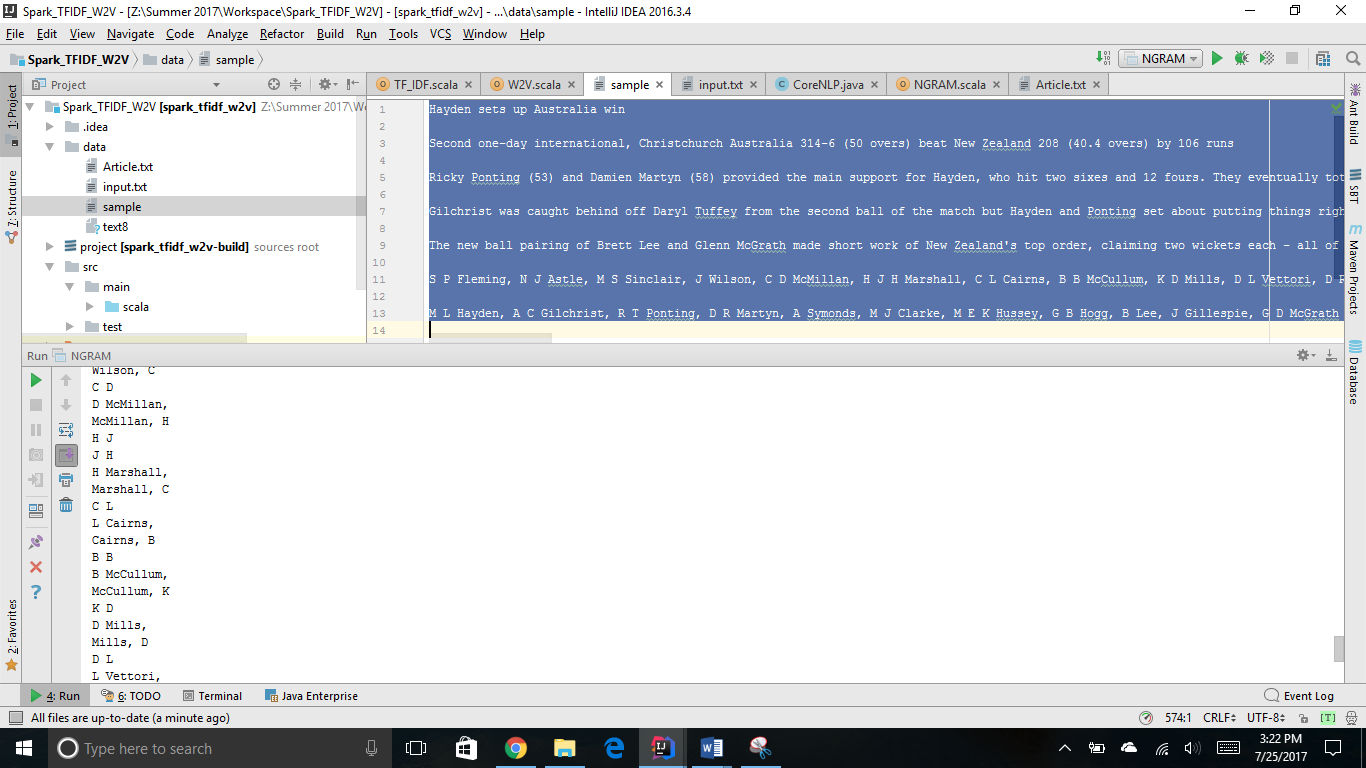


* 1. Try with Lemmatization





* 1. Try with NGrams



Compare the results from (a) , (b) and (c):

By comparing the outputs from W2V without NLP, with lemmatization and N-gram we can observe that the words are filtered after using the top TF-IDF words as input and after implementing these the ouputs for the questions are much more accurate than previous i.e without TF-IDF or N-gram.

1. ***Take home Question***

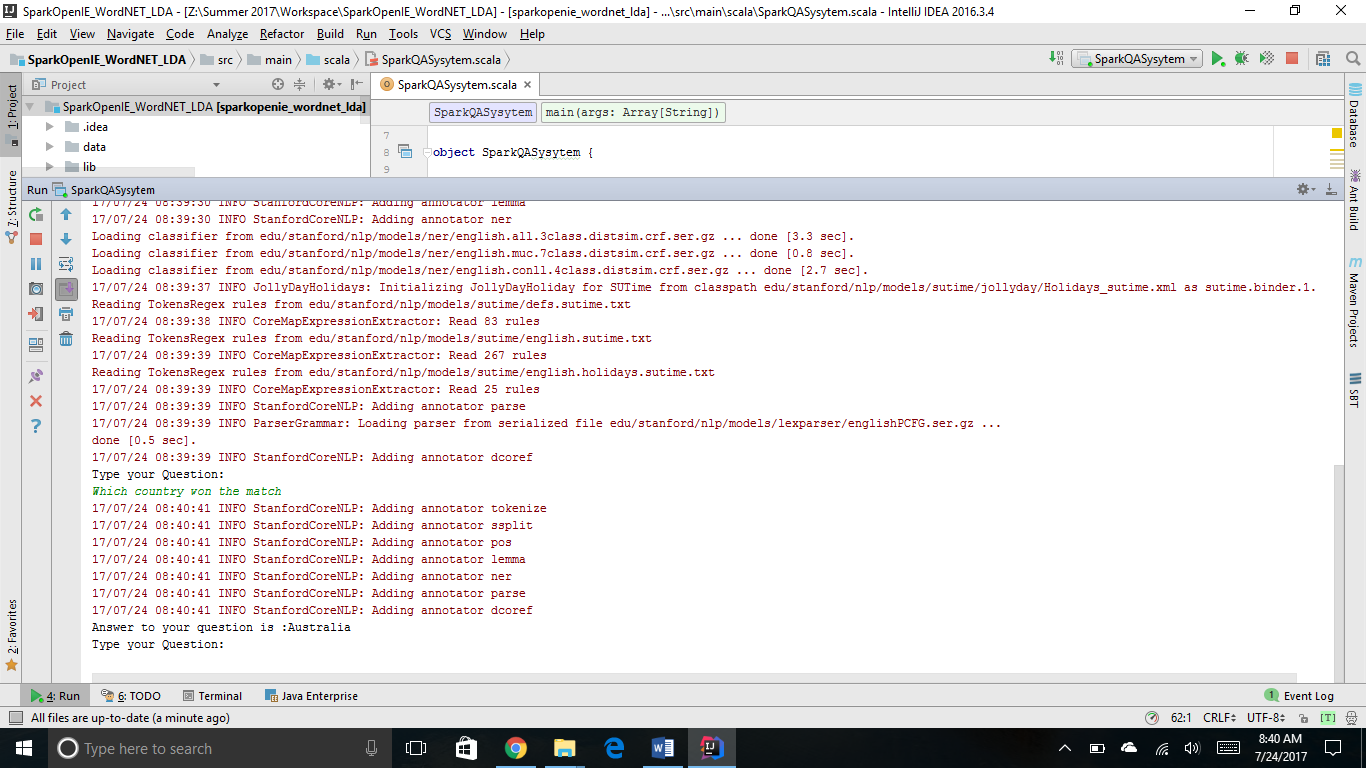
**Create a simple question answering system as an extension of the dataset and tasks done in (2). Continuation from Tutorial 2.**

**Question Answering system should be able to enrich the questions and answers based TF-IDF, W2V and N-Gram approaches. Use the diagram to guide you.**

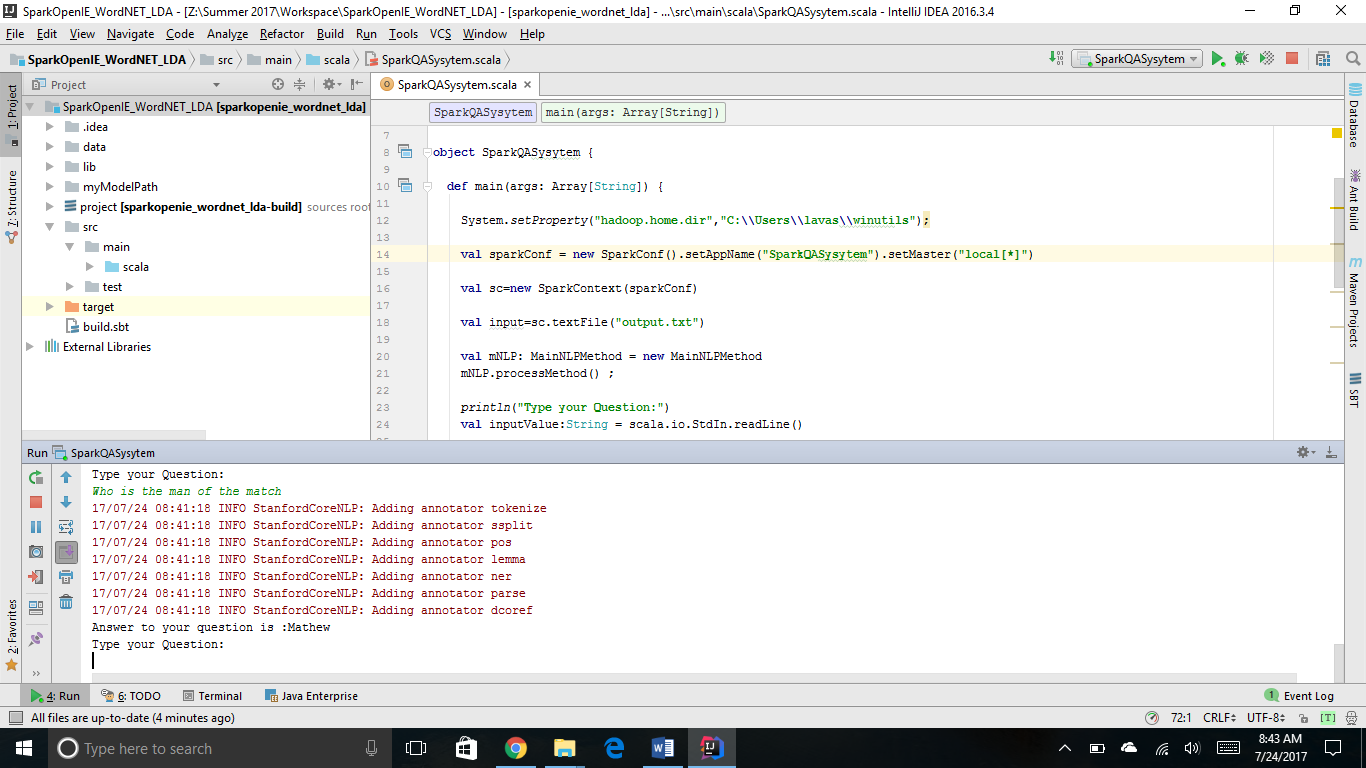
* 1. Use the Question Type

(https://en.wikipedia.org/wiki/Interrogative\_word). Suggestions are When, Who, What. Also try your questions with Google Search engine (<https://google.com>) or the Yahoo AnswerDemo (<https://answers.yahoo.com/>). **See additional details based Tutorial 3 PPT**

**Question 1:**



**Question 2:**



**Question 3:**

