PAGE CLASSIFIER

2ND TEAM PROJECT

2ND of November, 2016

THE MAIN IDEA

- We are introducing a program that can take a text of any size, read it, analyze it and output number of top keywords, which are the most repeated words in that text and can be most likely describing that text.
- This idea is a necessary step to make a web crawler for building a search engine, because keywords are always used during searching, along with another factors, to quickly find effective results.

FUNCTIONS

- Text_Cleaner(txt)
- get_base_form(word)
- CorrectWord(word)
- mergeSort(alist)
- Counter(alist)
- get_Max(Num,LastMax)

Text_Cleaner(txt)

This function receives the whole text, then starts to search for any special character within the words to remove it.

Then, it calls the get_base_form function, which checks if the word is a known verb, and determine it's tense, then return it to it's first base form.

get_base_form(word)

This function receives the a single word (string) from Text_Cleaner function.

First, it calls CorrectWord function to get the range which this function will search for the verb within.

Finally, it searches for the verb, and return the base form of it, if it was found.

mergeSort(alist)

This function receives the whole corrected text (list of strings), then it sorts all the words (strings) alphabetically, to make it easier to count the duplicated words later in the Counter function.

Counter(alist)

This function receives the whole corrected and sorted text, then it returns it in a new list, which is a list of lists, each list has two indexes, the first index contains the word, and the second index contains the number of it's clones.

get_Max(Num,LastMax)

This a recursive function, it takes the new list that Counter function has returned, then it finds the most repeated word and prints it if not in common words.

The function calls it self again to find the next most repeated word and print it, until it finishes displaying the number of top keywords the user wanted, it returns to the first function and stops.

SIMPLE EXAMPLE

Input = " Ahmed is an engineering student, Ahmed's skills are good enough for engineering."

myfile=['ahmed','is','an','engineering','student,','ahmed's','skills','are','good','enough','for','engineering.']

Text_Cleaner (myfile) + mergeSort(myfile):

→ ['ahmed','ahmed','an','are','engineering','engineering','enough','for','good','is','skills','student']

Counter(myfile)

→ [['ahmed' , 2] , ['an',1],['are',1],['engineering' , 2] , ['enough' , 1] , ['good' , 1] , ['skills' , 1] , ['student' , 1] ..]

get_Max(2 , len(myfile))

→ Output = ['ahmed' ,' engineering] : 2

['enough' ,'good' ,'skills' ,'student'] : 1

TEST PAGE 1 : UNITED STATES, WIKIPEDIA

United States

From Wikipedia, the free encyclopedia

Coordinates: 40°N 100°W

"United States of America", "America", "US", "U.S.", "USA", and "U.S.A." redirect here. For the landmass encompassing North and South America, see Americas. For other uses, see America (disambiguation), US (disambiguation), USA (disambiguation), and United States (disambiguation).

The **United States of America** (**USA**), commonly referred to as the **United States** (**U.S.**) or **America**, is a country composed of 50 states, a federal district, five major self-governing territories, and various possessions.[fn 1] The 48 contiguous states and federal district are in central North America between Canada and Mexico, with the state of Alaska in the northwestern part of North America and the state of Hawaii comprising an archipelago in the mid-Pacific. The territories are scattered about the Pacific Ocean and the Caribbean Sea. At 3.8 million square miles (9.8 million km^2)[18] and with over 324 million people, the United States is the world's third-largest country by total area (and fourth-largest by land area)[fn 2] and the third-most populous. It is one of the world's most ethnically diverse and multicultural nations, the product of large-scale immigration from many other countries.[24] The country's capital is Washington, D.C. and its largest city is New York City; other major metropolitan areas include Los Angeles, Chicago, Dallas, San Francisco, Boston, Philadelphia, Houston, Atlanta, and Miami. The geography, climate and wildlife of the country are extremely diverse.[25]

Paleo-Indians migrated from Asia to the North American mainland at least 15,000 years ago,[26] with European colonization beginning in the 16th century. The United States emerged from 13 British colonies along the East Coast. Numerous disputes between Great Britain and the colonies in the aftermath of the Seven Years War led to the American Revolution, which began in 1775. On July 4, 1776, as the colonies were fighting Great Britain in the American Revolutionary War, delegates from the 13 colonies unanimously adopted the Declaration of Independence. The war ended in 1783 with recognition of the independence of the United States by Great Britain, and was the first successful war of independence against a European colonial empire.[27] The current constitution was adopted in 1788, after the Articles of Confederation, adopted in 1781, were felt to have provided inadequate federal powers. The first ten amendments, collectively named the Bill of Rights, were ratified in 1791 and designed to guarantee many fundamental civil liberties.

The United States embarked on a vigorous expansion across North America throughout the 19th century,[28] displacing American Indian tribes, acquiring new territories, and gradually admitting new states until it spanned the continent by 1848.[28] During the second half of the 19th century, the American Civil War led to the end of legal slavery in the country.[29][30] By the end of that century, the United States extended into the Pacific Ocean,[31] and its economy, driven in large part by the Industrial Revolution, began to soar.[32] The Spanish–American War and World War I confirmed the country's status as a global military power. The United States emerged from World War II as a global superpower, the first country to develop nuclear weapons, the only country to use them in warfare, and a permanent member of the United Nations Security Council. The end of the Cold War and the dissolution of the Soviet Union in 1991 left the United States as the world's sole superpower.[33]

The United States is a highly developed country, with the world's largest economy by nominal GDP. It ranks highly in several measures of socioeconomic

United States of America

Flag

Great Seal

Motto:
"In God we trust"[1][2]
Other traditional mottos [show]

Anthem: "The Star-Spangled Banner"

0:00 CC MENU

March: "The Stars and Stripes Forever"[3]

0:00 CC MENU

TEST PAGE 1 : UNITED STATES, WIKIPEDIA

```
Words found:  36173

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.3749673366546631 seconds
mergeSorter:      0.2968747615814209 seconds
Counter:          0.015609025955200195 seconds

Top Keywords:
['jump'] : 620
['state'] : 450
['retrieved'] : 447
['united'] : 366
['u.s'] : 223
['american'] : 204
['isbn'] : 177
['p'] : 172
['world'] : 136
['october'] : 120

get_Max:          0.078125 seconds
Total Time:       0.781205415725708 seconds
>>>
```

- our program -

Top Keyword Density

Top 10 | Exclude grammar words ON

1 Word | 2 Words | 3 Words

1. up	638 *(1.7%)*
2. jump	620 *(1.6%)*
3. retrieved	447 *(1.2%)*
4. states	408 *(1.1%)*
5. united	386 *(1%)*
6. american	211 *(0.5%)*
7. isbn	177 *(0.5%)*
8. 2015	176 *(0.5%)*
9. 2013	172 *(0.4%)*
10. 2014	162 *(0.4%)*

- wordcounttools.com -

TEST PAGE 2 : EGYPT, WIKIPEDIA

Egypt

From Wikipedia, the free encyclopedia

Coordinates: 26°N 30°E

This article is about the modern country. For the ancient realm, see Ancient Egypt.
For other uses, see Egypt (disambiguation).

Egypt (/ˈiːdʒɪpt/ ***EE-jipt***; Arabic: مِصر *Miṣr*, Egyptian Arabic: مَصر *Maṣr*, Coptic: Ⲭⲏⲙⲓ *Khemi*), officially the **Arab Republic of Egypt**, is a transcontinental country spanning the northeast corner of Africa and southwest corner of Asia by a land bridge formed by the Sinai Peninsula. Egypt is a Mediterranean country bordered by the Gaza Strip and Israel to the northeast, the Gulf of Aqaba to the east, the Red Sea to the east and south, Sudan to the south, and Libya to the west. Across the Gulf of Aqaba lies Jordan, and across from the Sinai Peninsula lies Saudi Arabia, although Jordan and Saudi Arabia do not share a land border with Egypt. It is the world's only contiguous Afrasian nation.

Egypt has among the longest histories of any modern country, emerging as one of the world's first nation states in the tenth millennium BC.[14] Considered a cradle of civilisation, Ancient Egypt experienced some of the earliest developments of writing, agriculture, urbanisation, organised religion and central government. Iconic monuments such as the Giza Necropolis and its Great Sphinx, as well the ruins of Memphis, Thebes, Karnak, and the Valley of the Kings, reflect this legacy and remain a significant focus of archaeological study and popular interest worldwide. Egypt's rich cultural heritage is an integral part of its national identity, which has endured, and at times assimilated, various foreign influences, including Greek, Persian, Roman, Arab, Ottoman, and European. One of the earliest centres of Christianity, Egypt was Islamised in the seventh century and remains a predominantly Muslim country, albeit with a significant Christian minority.

With over 90 million inhabitants, Egypt is the most populous country in North Africa and the Arab world, the third-most populous in Africa (after Nigeria and Ethiopia), and the fifteenth-most populous in the world. The great majority of its people live near the banks of the Nile River, an area of about 40,000 square kilometres (15,000 sq mi), where the only arable land is found. The large regions of the Sahara desert, which constitute most of Egypt's territory, are sparsely inhabited. About half of Egypt's residents live in urban areas, with most spread across the densely populated centres of greater Cairo, Alexandria and other major cities in the Nile Delta.

Modern Egypt is considered to be a regional and middle power, with significant cultural, political, and military influence in North Africa, the Middle East and the Muslim world.[15] Egypt's economy is one of the largest and most diversified in the Middle East, and is projected to become one of the largest in the 21st century. Egypt is a member of the United Nations, Non-Aligned Movement, Arab League, African Union, and Organisation of Islamic Cooperation.

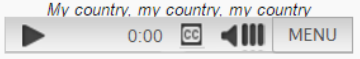
Contents [hide]

Arab Republic of Egypt
جمهورية مصر العربية [show]

Flag

Coat of arms

Anthem: *Bilady, Bilady, Bilady*
My country, my country, my country



TEST PAGE 2 : EGYPT, WIKIPEDIA

```
Words found:  21304

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.21875333786010742 seconds
mergeSorter:      0.17181968688964844 seconds
Counter:          0.01564478874206543 seconds

Top Keywords:
['egypt'] : 415
['jump'] : 265
['egyptian', 'retrieved'] : 157
['world'] : 70
['february'] : 61
['new', 'state'] : 53
['cairo', 'country', 'main'] : 49
['august'] : 47
['president'] : 46
['july'] : 44

get_Max:          0.07812380790710449 seconds
Total Time:       0.5156481266021729 seconds
>>>
```

- our program -

Top Keyword Density

Top 10 | Exclude grammar words ON

1 Word | 2 Words | 3 Words

1. egypt	406 *(1.8%)*
2. up	272 *(1.2%)*
3. jump	265 *(1.2%)*
4. egyptian	176 *(0.8%)*
5. retrieved	157 *(0.7%)*
6. 2013	83 *(0.4%)*
7. 2014	83 *(0.4%)*
8. egypt's	79 *(0.4%)*
9. world	68 *(0.3%)*
10. february	61 *(0.3%)*

- wordcounttools.com -

TEST PAGE 3 : SCIENCE ARTICLE, DISCOVER

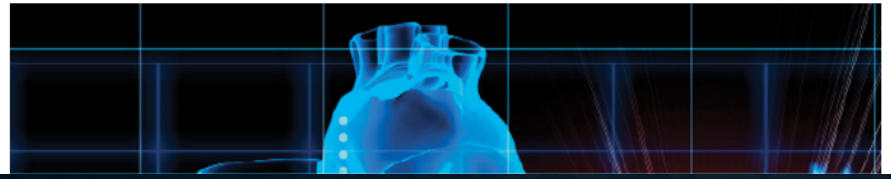
FROM THE DECEMBER 2014 ISSUE

Solving Biology's Mysteries Using Quantum Mechanics

The new field of quantum biology applies the craziness of quantum physics to biology's most fundamental processes.

By Zeeya Merali | Monday, December 29, 2014

RELATED TAGS: BIOTECHNOLOGY, BIOLOGY



NEW ON DISCOVER

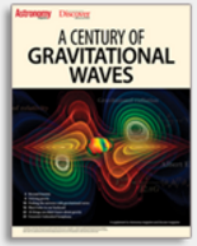
Ever Feel Like You're Being Watched?

The Lowly Leech

Arctic sea ice extent is trending at record low levels

October has been dramatically warm and dry in much of the United States — and relief is not yet in sight

Scientists snare their first ever observations of a solar wave



TEST PAGE 3 : SCIENCE ARTICLE, DISCOVER

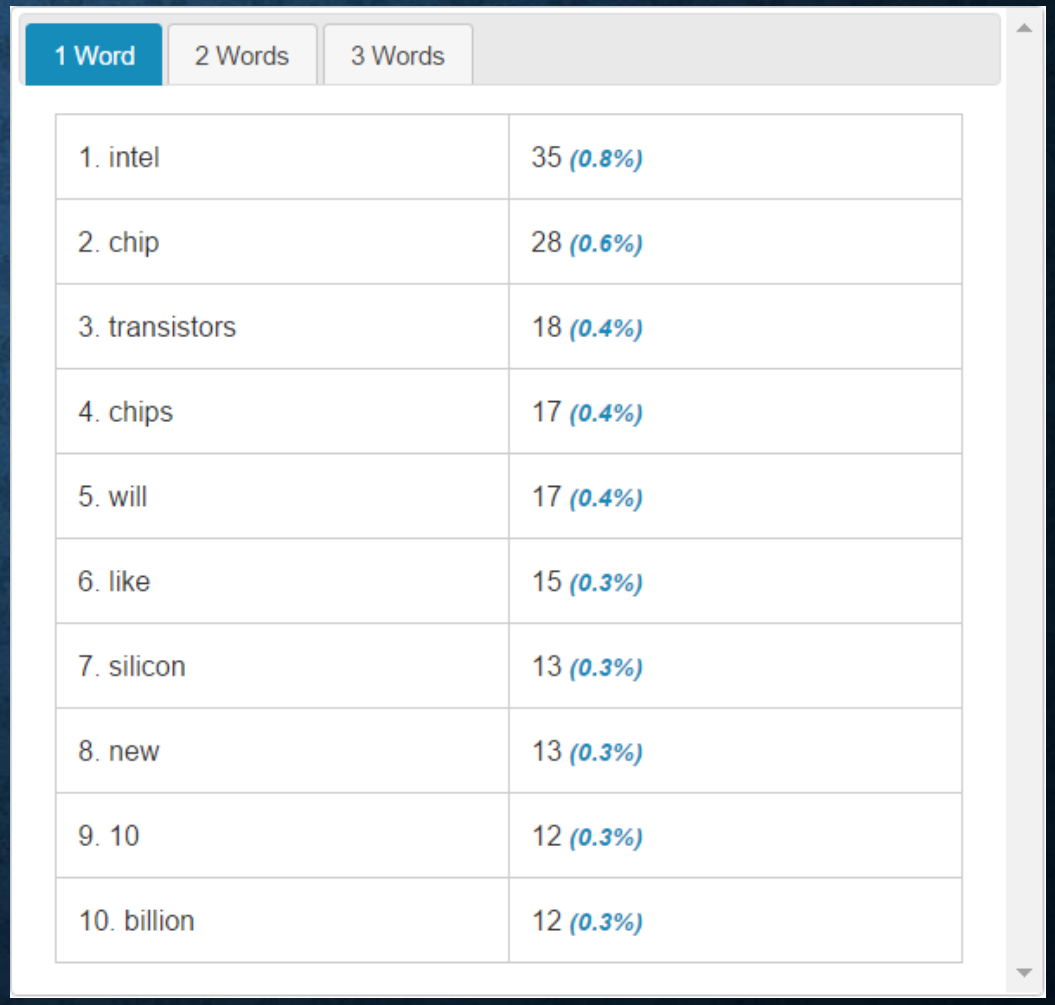
```
Words found:  4362

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.0468592643737793 seconds
mergeSorter:      0.03126835823059082 seconds
Counter:          0.0 seconds

Top Keywords:
['intel'] : 47
['chip'] : 46
['transistors'] : 16
['make'] : 14
['new', 'silicon', 'take', 'use', 'work'] : 13
['billion', 'design', 'one'] : 11
['circuits', 'company', 'each'] : 10
['world'] : 9
['bloomberg', 'create', 'data', 'need', 'run', 'year'] : 8
['add', 'call', 'designers', 'e5', 'half', 'layers', 'mask',
rcent', 'those'] : 7

get_Max:          0.062486171722412 11 seconds
Total Time:       0.1562352180480957 seconds
>>>
```



1 Word | 2 Words | 3 Words

1. intel	35 *(0.8%)*
2. chip	28 *(0.6%)*
3. transistors	18 *(0.4%)*
4. chips	17 *(0.4%)*
5. will	17 *(0.4%)*
6. like	15 *(0.3%)*
7. silicon	13 *(0.3%)*
8. new	13 *(0.3%)*
9. 10	12 *(0.3%)*
10. billion	12 *(0.3%)*

- our program -

- wordcounttools.com -

TEST PAGE 4 : TECHNOLOGY ARTICLE, BLOOMBERG

How Intel Makes a Chip

The development of a microprocessor is one of the riskiest, costliest, and most technically complex feats in business.

by Max Chafkin, Ian King

from BloombergBusinessweek

June 9, 2016 — 11:02 AM EET

TEST PAGE 4 : TECHNOLOGY ARTICLE, BLOOMBERG

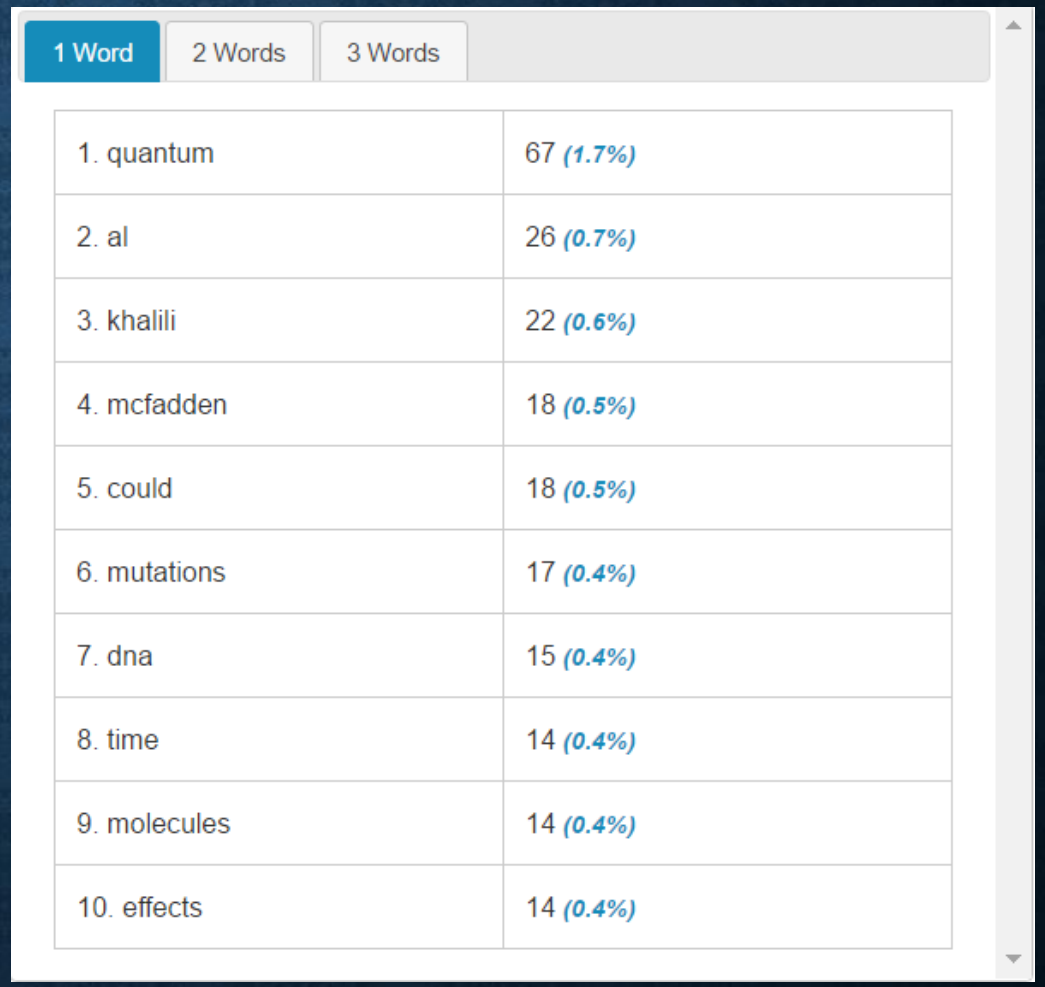
```
Words found:  3900

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.0468297004699707 seconds
mergeSorter:      0.0312497615814209 seconds
Counter:          0.0 seconds

Top Keywords:
['quantum'] : 64
['al-khalili'] : 25
['mcfadden'] : 20
['biology', 'dna', 'mutations'] : 16
['effects', 'molecules', 'out'] : 14
['biological', 'time'] : 13
['energy', 'experiments'] : 12
['take'] : 11
['atom', 'explain', 'farrow', 'hydrogen', 'physicists',
ersity', 'use', 'years'] : 9
['both', 'discover', 'find', 'first', 'make', 'mutation',
rposition', 'work'] : 8

get_Max:          0.06250452995300293 seconds
Total Time:       0.1562061309814453 seconds
>>>
```



1 Word | 2 Words | 3 Words

1. quantum	67 *(1.7%)*
2. al	26 *(0.7%)*
3. khalili	22 *(0.6%)*
4. mcfadden	18 *(0.5%)*
5. could	18 *(0.5%)*
6. mutations	17 *(0.4%)*
7. dna	15 *(0.4%)*
8. time	14 *(0.4%)*
9. molecules	14 *(0.4%)*
10. effects	14 *(0.4%)*

- our program -

- wordcounttools.com -

TEST PAGE 5: FC BARCELONA ARTICLE

FCB SITES MEMBERS PENYES 18.8M 94.6M

FCBARCELONA

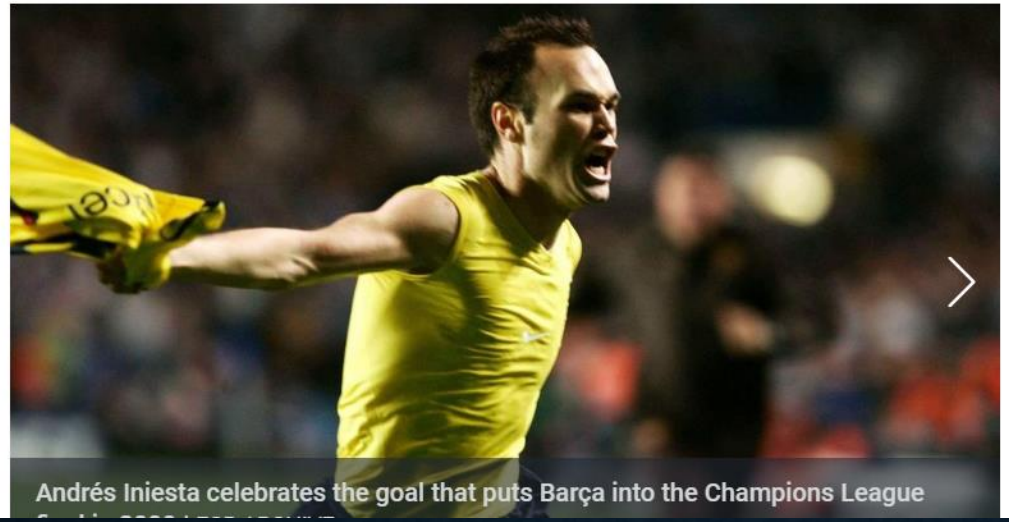
NEWS FOOTBALL SPORTS CLUB CAMP NOU FANS TICKETS VIP TOUR SHOP GAMEPASS

Home › Club › History : 2008-16. The best years in our history

Share

2008-16. The best years in our history

Barça delight the world with their brand of football that is appreciated by football fans from all over the world. They win three more Champions League titles in a decade of dominance for the blaugranes



Andrés Iniesta celebrates the goal that puts Barça into the Champions League

CLUB SCHEDULE

VIEW

TEST PAGE 5: FC BARCELONA ARTICLE

```
Words found:  1632

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.015581607818603516 seconds
mergeSorter:      0.0 seconds
Counter:          0.0 seconds

Top Keywords:
['win'] : 26
['barça'] : 24
['barcelona'] : 17
['fc', 'league'] : 16
['champions', 'cup', 'world'] : 13
['fcb', 'team'] : 12
['archive', 'club'] : 11
['history', 'three'] : 10
['football', 'guardiola'] : 9
['first', 'two'] : 8

get_Max:          0.0625004768371582 seconds
Total Time:       0.10931658744812012 seconds
>>>
```

1 Word | 2 Words | 3 Words

1. barça	22 *(1.3%)*
2. barcelona	18 *(1.1%)*
3. league	16 *(1%)*
4. fc	16 *(1%)*
5. won	14 *(0.8%)*
6. champions	13 *(0.8%)*
7. fcb	13 *(0.8%)*
8. team	12 *(0.7%)*
9. world	12 *(0.7%)*
10. archive	11 *(0.7%)*

- our program -

- wordcounttools.com -

TEST PAGE 6 : MEDIA ARTICLE, BBC

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

culture

Home | Film | Music | Art | Style | Books | Design | Columns | Programmes | Calendar

Music

What makes a great horror movie soundtrack?

A great score for a horror film can be as terrifying as the pictures - but what is it about the best ones that makes us feel so frightened? Arwa Haider looks for answers.

Related Stories

TEST PAGE 6 : MEDIA ARTICLE, BBC

```
Words found:  1722

Show (n) of top keywords. Insert (n): 10

Text_Cleaner:     0.015586614608764648 seconds
mergeSorter:      0.0 seconds
Counter:          0.0 seconds

Top Keywords:
['horror'] : 18
['music'] : 15
['bbc'] : 14
['share'] : 13
['films'] : 12
['culture'] : 9
['frizzi', 'score'] : 8
['carpenter', 'fear', 'film', 'movies'] : 7
['alamy', 'halloween', 'sound', 'soundtrack', 'things']
['create', 'images', 'include', 'make', 'new', 'out', 's
 'time'] : 5

get_Max:          0.06249499320983887 seconds
Total Time:       0.1093285083770752 seconds
>>>
```

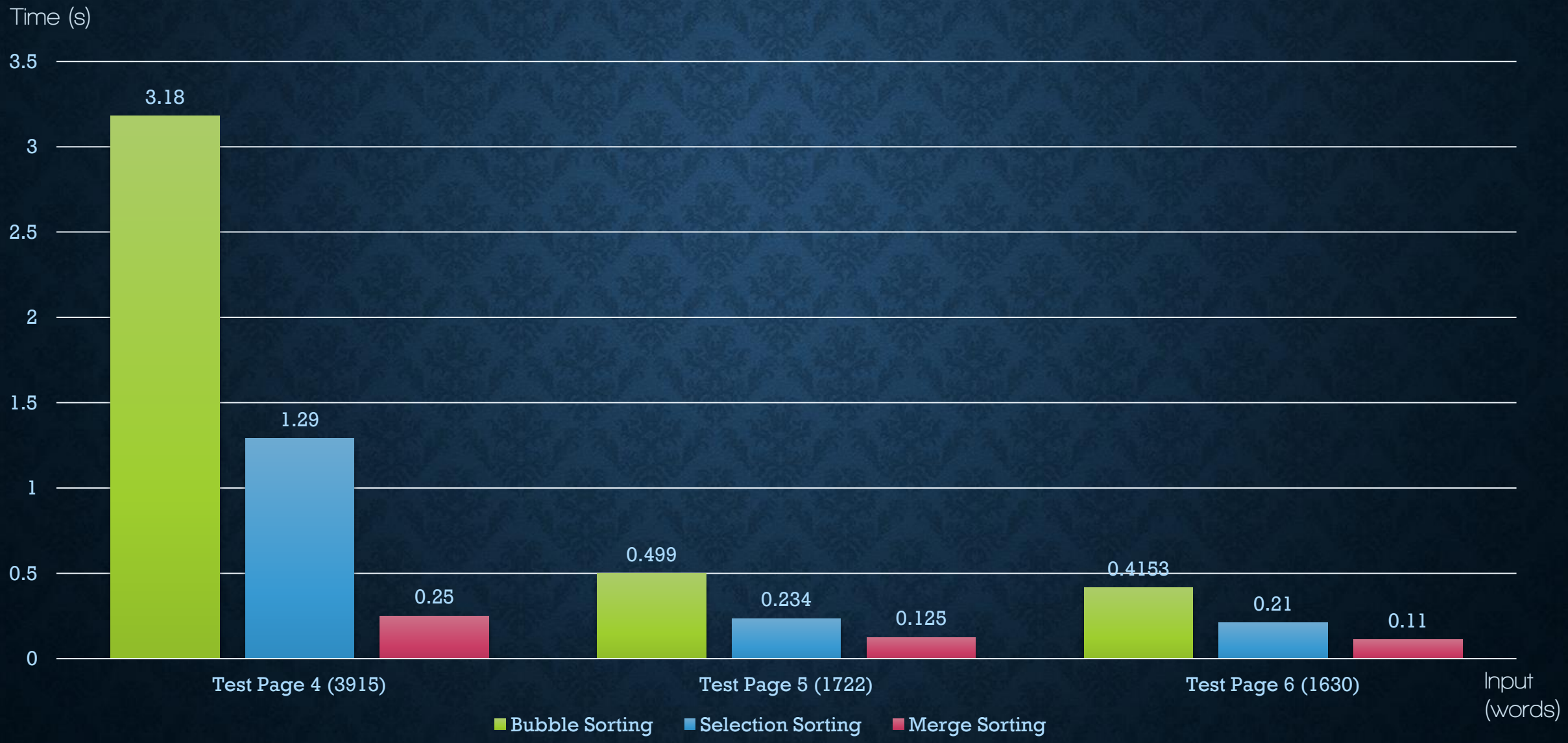
- our program -

1 Word | 2 Words | 3 Words

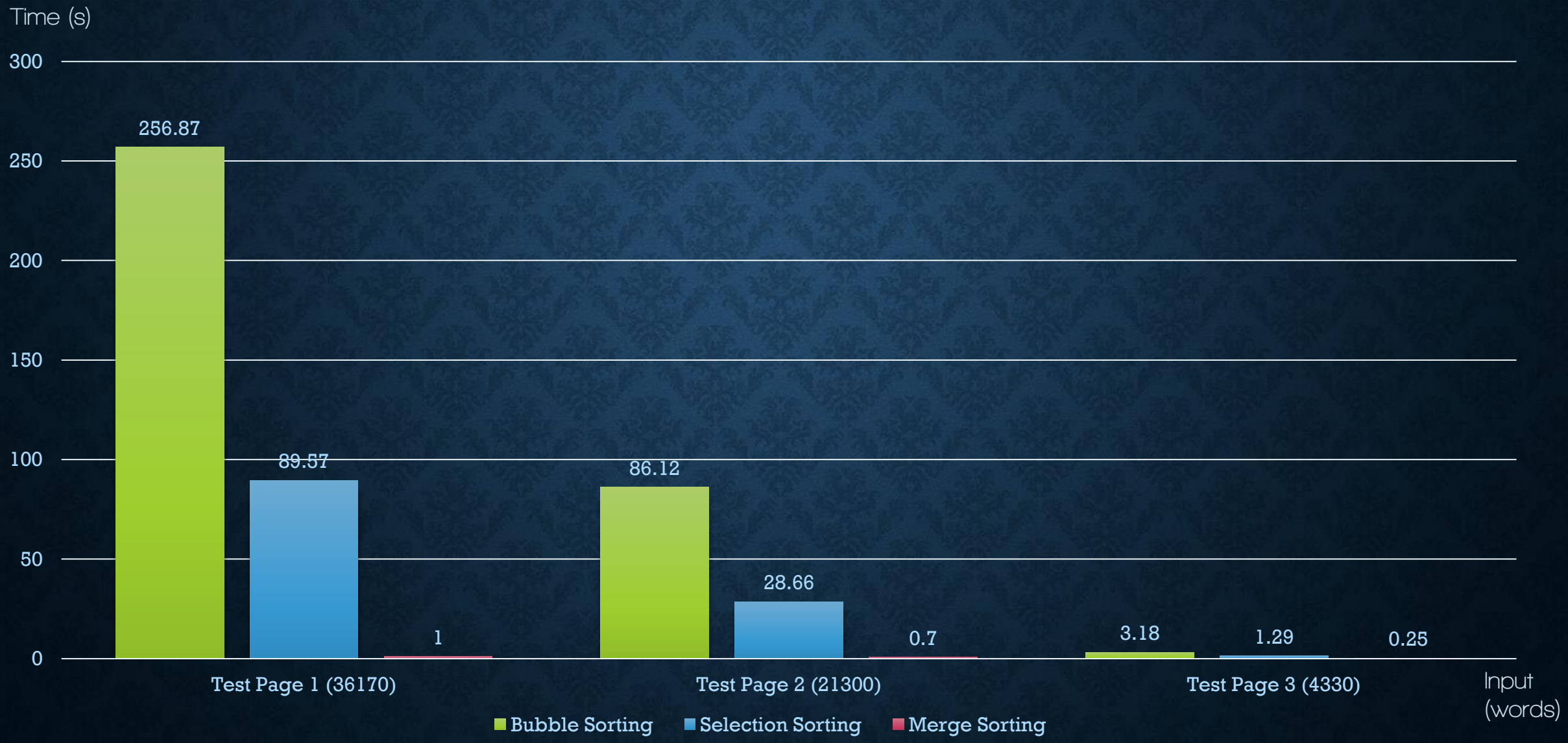
1. credit:	21 *(1.2%)*
2. horror	18 *(1%)*
3. bbc	15 *(0.9%)*
4. music	15 *(0.9%)*
5. share	13 *(0.7%)*
6. films	12 *(0.7%)*
7. culture	9 *(0.5%)*
8. score	8 *(0.5%)*
9. frizzi	7 *(0.4%)*
10. fear	7 *(0.4%)*

- wordcounttools.com -

COMPARING BETWEEN THE SORTING ALGORITHMS – LOW INPUTS



COMPARING BETWEEN THE SORTING ALGORITHMS – HIGH INPUTS



NOTE: At 578768 words, the merge sorting function needed only 8.4s to sort, while the rest of the program needed additional 8.12s to finish. Total about 16s.

THANK YOU FOR LISTENING !

Abdulrahman Hamdy (L)

Amr Khaled

Zeyad Elsawy

Samir Mohamed

Sherif Ezzat

Abdulrahman Tarek

Ahmed Waleed

Khaled Al Deeb

Abdulrahman Emam

Mohamed Lebda

"This is not about who is better, it's about getting better."

Let's keep learning.