FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

DEPARTMENT OF INFORMATION AND MEDIA TECHNOLOGY

SECOND SEMESTER 2015/2016 EXAMINATION

COURSE CODE:	IMT524
COURSE TITLE:	INFORMATION RETRIEVAL SYSTEMS
CREDIT UNITS:	3 UNITS
TIME ALLOWED:	**2hrs 30min**
COURSE LECTURER(S):	Dr. I. O. Oyefolahan and Mrs Stella Etuk
NUMBER OF QUESTIONS:	4
NUMBER OF PAGES:	2 (INCLUDING THIS PAGE)

INSTRUCTIONS

- Answer all questions
- Do **not** use red pen
- Please use a clear handwriting
- This exam is closed book, closed notes, closed laptop and closed cell phone
- Please use non-programmable calculators only

Question 1.

a. Write short note on the following as related to Information Retrieval (IR)
   i. Tokenization
   ii. Document Parsing
   iii. Stopping
   iv. Stemming

(4mks)

b. Explain the vocabulary mismatch problem in IR (3mks)

c. Search Engine architectural components supports 2 major functions: Index processing and query processing. Explain the query processing. (8mks)

Question 2.

a. Carefully analyze the process of retrieving web pages from the web servers

(5mks)

b. Explain two (2) challenges of web crawling and how the crawler does attempts to tackle the challenges. (5mks)

c. Duplicate and near-duplicate documents occur in many situations on the web. What algorithm or technique has been put in place to detect duplicate document? Explain the technique. (5mks)

Question 3.

a. "One of the most obvious features of text from a statistical point of view is that the distribution of word frequency is much ***skewed***. Discuss. (3mks)

b. 'To build search engines that search web pages, you first need a copy of the pages that you want to search. Unlike some of the other sources of text, web pages are easily copied, since the pages are meant to be retrieved over the Internet by browsers. This instantly solves one of the major problems of getting information to search, which is how to get the data from the place they are stored to the search engines':

   Considering the word collection given above as a document, show in a tabular form using Zipf's law the first 5 most common words in the document together with their frequency (f), rank (r), probability of occurrence (Pr) and its ($r.Pr$) values.

(10mks)

c. From the table generated in 3b above, calculate the number of words with the same frequency given that a word that occur n times has a rank r_n (2mks)

Question 4

The index terms given below are derived from the following three descriptions about corruption from investopedia, transparency international and wikipedia respectively:

"Corruption is dishonest behavior by those in positions of power, such as managers or government officials. Corruption can include giving or accepting bribes or inappropriate gifts, double dealing, under-the-table transactions, manipulating elections, diverting funds, laundering money and defrauding investors."

"Corruption is the abuse of entrusted power for private gain. It can be classified as grand, petty and political, depending on the amounts of money lost and the sector where it occurs."

"Corruption is a form of dishonest or unethical conduct by a person entrusted with a position of authority, often to acquire personal benefit. Corruption may include many activities including bribery and embezzlement, though it may also involve practices that are legal in many countries."

{abuse, accepting, acquire, acts, amounts, an, and, as, authority, be, behavior, benefit, bribery, bribes, can, capacity, conduct, corruption, countries, dealing, defrauding, depending, dishonest, diverting, double, embezzlement, entrusted, form, funds, gain, gifts, giving, government, grand, in, include, including, is, it, involve, laundering, legal, lost, managers, manipulating, many, money, of, officials, often, on, or, person, petty, political, position, positions, power, practices, sector, such, table, that, the, thought, unethical, under.}

In information retrieval, indexes are designed to support search, and inverted index has been mentioned as the most common data structure to make searching faster. Based on the information above, answer the following questions:

a. Describe briefly what make index terms to be described as "inverted" (2mks)
b. Perform simple inverted index with postings (5mks)
c. Inverted index with counts is known for supporting better ranking algorithm, use the indexes and the sentences above to perform an inverted index with counts (4mks)
d. Perform an inverted index with positions as an approach to supporting proximity matches (4mks)