Name: Mrinal Kadam
ID: 3135945534

Q1 1) F-measure
2) Discounted Cumulative Gain

Q2
http://www.betasoftware.com/~myfolder
http://www.betasoftware.com:8081/a±b
http://www.betasoftware.com/FOLDERX/marie-smith

Q3 1) DNS caching
2) Pre-Fetching Client

Q4 n-k+1

Q5 NO

Q6 JS(D1, D2) = 0.33
JS(D1, D3) = 0
JS(D1, D4) = 0.13 (after rounding off)

Q7 Result by Church and Gale that was cited which is similar to Heaps Law-

$|V| > O(N\ 1/2)$
where N : number of tokens &
V : vocabulary set

The size of the vocabulary set changes with the number of tokens.

Q8 Part 1: takes every non-alphabetic character from 'shakes.txt' and translates it into a newline character
Part 2: sorts all the words in each line
Part 3: takes the sorted file and gives the count of each unique word
Part 4: sorts numerically by descending order of count

Q9 $|A \cup B| = |A| + |B| - |A \cap B|$

Q10

	R3	**NR4**	**R2**	**R1**	**NR1**	**NR2**	**R4**	**R5**	**NR3**	**NR5**
Recall	0.2	0.2	0.4	0.6	0.6	0.6	0.8	1	1	1
Precision	1	0.5	0.66	0.75	0.6	0.5	0.57	0.62	0.55	0.5

Q11 1) Resnik method
2) Dekang Lin method

Q12 The Lesk method compares the glosses of 2 words and the more the similarity in the glosses, the higher will be similarity of the words.

Q13 Accuracy:
(tp + tn) / (tp + fp + fn + tn)

Q14 Politeness: Back Queues
Prioritization: Front Queues

Q15 50

Q16 Website's Root Directory

Q17 P(n) = k*((n)^-1)
so, inversely proportional to the rank

On a log-log plot, power laws give a straight line with slope c.
The x axis will have log of result-rank.
The y axis will have log of frequency.

Therefore, Axes: -
y-axis: log(P(n)) &
x-axis: log(n)
for the graph to be a straight line.

Q18 Mumbai - M510
Bombay - B510
Resulting numbers are not the same.

Q19 compute, computes, computing, computed

Q20 Stemming identifies the common root form of a word by removing or replacing word suffixes (e.g. flooding is stemmed as flood), while lemmatization identifies the infected forms of a word and returns its base form (e.g. better is lemmatized as good).

Q21 Bi-word indexing expands the dictionary.
Positional indexing expands the postings list.

Q22 64

Q23 Youtube created a fingerprint database of copyrighted content called contentID to monetize it's websites.

Q24 1) Computationally difficult to compute an alphanumeric text that has a given hash.
2) Small change to text yields a totally different value
3) Extremely unlikely that two slightly different messages will have same hash.

Q25 b) -24

Q26 b) 33.33%

Q27 3) 0

Q28 shouldVisit()

Q29 shouldVisit():
This function should be implemented to specify whether the given URL should be crawled or not (based on the crawling logic).

visit():
This function is called when a page is fetched and ready to be processed by your program.

Q30 1) crawlStorageFolder
2) numberOfCrawlers
3) maxPagesToFetch

Q31 shouldVisit() - boolean
visit() - no return type since void

Q32

	A	B	C	D	E	F
A	0	0	0	1/4	1	1/3
B	1/3	0	0	0	0	1/3
C	1/3	1/4	0	1/4	0	0
D	0	1/4	1/2	0	0	0
E	0	1/4	1/2	1/4	0	1/3
F	1/3	1/4	0	1/4	0	0

Q33

Node 1	**5/36**
Node 2	**13/36**
Node 3	**5/36**
Node 4	**1/4**
Node 5	**1/18**
Node 6	**1/18**

Q34 T1 , T3, T2