Assignment 2 Submission Sheet

TASK 1: Warm Up

1. Number of records: **9999**
2. Number of attributes: **500**

TASK 2: Frequent Itemsets

1. What is the most frequent item (i.e., single keyword): **of**
2. In how many queries does it occur: **960 times**
3. Which support value corresponds to a support count of 100 records? **0.01**
4. How many frequent itemsets have you found with this min-support? **125**
5. What is the maximum size of your frequent itemsets? **24**

TASK 3: Impact of the Support Parameter

Minimum Support	Number of Frequent Items
0.001	715
0.002	394
0.003	224
0.004	139

Which value of min-support leads to the discovery of about 150 itemsets? **0.0038**

TASK 4: Rule Generation

1. How many rules have you identified with min-confidence .8? **6**
2. Indicate a high-confidence rule X -> Y where Y -> X has lower confidence.
   York -> New has confidence of 0.897. New -> York has confidence of 0.397
3. Why is the confidence of Y -> X lower than the confidence of X -> Y?
   Because when someone has a search for ‘York’ they probably search for ‘New York’ since this is the US where the data originates. Very rarely would someone search for the English city. Meanwhile ‘New’ could be used for many other things, for example ‘New Vacuum’.

TASK 5: Impact of confidence

Minimum confidence	**Number of Rules**
0.8	6
0.6	9
0.4	13
0.2	21
0.1	32

TASK 6: Rule Interpretation

1. **Some interesting rules:**

Estate -> Real	**Conf. 0.917**	**Lift: 129.662**
(To, A) -> How	**Conf. 0.519**	**Lift: 28.808**
Pictures -> Of	**Conf. 0.457**	**Lift: 4.755**

2. **Some unexpected rules:**

Is -> What	**Conf. 0.446**	**Lift: 35.424**
The -> Of	**Conf. 0.234**	**Lift: 2.439**
For -> Sale	**Conf. 0.153**	**Lift: 15.538**