Ethan Hadley

1:

- select minimum frequency and confidence threshold.
- find all single items which appear more than the threshold frequency. toss the rest.
- generate all possible pairs of items (2-itemsets)
- count their frequency, throwing out the ones which are under the threshold.
- repeat for all possible 3-itemsets, 4-itemsets, etc.
- stop when no larger itemsets above the threshold can be found.
- use the sets to generate associative rules, that state a correlation exists between some subset of the itemset and the remaining subset.
- test all these rules for accuracy and keep the ones that pass the confidence threshold

2:

- The Apriori property is that all subsets of a common set are also common. Supopose we have a common set A, containing subset B. Any time A occurs, B occurs. But the converse is not necessarily true. So B can only occur more times in the dataset than A does. Meaning B, the subset, can only be as frequent or more frequent than any set it belongs to. We use this to prune candidates itemsets which have infrequent subsets. If the subset is infrequent, the parent set must be less frequent, so couldnt not pass the threshold.

3:

a:

- so the minimum support is 2 occurrences. That means our 1-itemsets are {I1, I2, I3, I4, I5}
- the possible 2-sets are: {{I1,I1}{I1,I2}{I1,I3}{I1,I4}{I1,I5}{I2,I1}{I2,I2}{I2,I3}{I2,I4}{I2,I5}{I3,I1}{I3,I2}{I3,I3}{I3,I4}{I3,I5}{I4,I1}{I4,I2}{I4,I3}{I4,I4}{I4,I5}{I5,I1}{I5,I2}{I5,I3}{I5,I4}{I5,I5}}
- The ones which occur twice or more: {{I1,I2},{I2,I4},{I2,I3},{I1,I3},{}}

4:

- Hash-Based techniques: as we traverse the dataset to count the frequent itemsets, we can hash itemsets into a buckets to count theirr frequency at the same time. Once we are done, we can prune any sets which contained an infrequent single item.

- Mark transactions once we know they contain no frequent subsets. Once we have seen that the transaction contains no frequent k-set, we know it cannot contain a frequent k+1-set. We can skip it when scanning the dataset for all future iterations.

- Sampling. If we assume some homogeneity in the dataset, we can simply run the algorithm on a subset of the dataset to improve speed. In most cases we will gather the same patterns as if we used the full dataset, unless our selection was biased in some way.