COSC-254 Data Mining — Covered Topics

Topics are split by chapter, and are presented in bullet-point lists.
Very important topics are in bold, not-so-important topics are in italic.
Everything else is “just” important.

MapReduce
- General framework: map → combiners → shuffle → reduce.
- **map Function:** input, output, execution.
- **reduce Function:** input, output, execution.
- Combiners: input, output, execution, sufficient condition for using a reduce function in a combiner
- **Computational and communication complexity.**

Pattern mining
- Motivation and applications
- Definitions: dataset, items, itemsets, support, frequency, search space, itemset lattice,
- **Frequent Itemsets:** definition of the task, anti-monotonicity property of the support, minimum support threshold, negative border.
- Association rules: confidence, definition of the task, sufficient condition for pruning ARs, algorithm for mining ARs
- **Apriori Algorithm:** general idea, algorithm, improving the I/O.
- **Eclat algorithm:** motivation, tidsets, prefix equivalence classes, comparison with Apriori.
- Maximal and Closed Frequent Itemsets: motivation, definition, reconstructing the FIs from the CFIs.

Data streams
- Motivation, applications, queries, approximations
- **Reservoir sampling:** motivation, algorithm, proof.
- **Bloom filters:** hash functions, set membership, bloom filter construction and analysis, union and intersection of bloom filters.
- **The Flajolet-Martin approach for distinct counting:** problem, algorithm, analysis.
- **The DGIM approach to counting on sliding window:** problem, algorithm, analysis.

---

1These topics are "not-so-important" for the purpose of this course and for preparing for the final. They are all of crucial importance in the field of algorithms for data mining and in the practice of data analysis at scale.

2If the name of the topic is in bold/italic/normal font, everything that follows is very important/not-so-important/“just” important, respectively, unless it appears in a different font.
Graph analysis

- Graphs: basic definitions, shortest paths, **BFS algorithm**
- Triangle counting: definitions, TRIÈST algorithms (no proofs), exact counting, **smart algorithm for exact counting (heavy hitters)**, counting triangles in MapReduce.
- **Social network analysis**: homophily, triadic closure, clustering coefficient, preferential attachment and its consequences.
- **Centrality measures**: degree centrality, BFS, closeness centrality, betweenness, exact and approximation algorithms for centralities.
- **Finding communities**: definitions, hubs and communities, cuts, finding communities with betweenness centrality
- Link prediction: task, **neighborhood-based measures** (common neighbors, Jaccard, Adamic-Adar).

Link analysis

- The Web: data and applications
- Crawling: motivation, naïve algorithm, **duplicate and near-duplicate identification**
- Indexing: task, building an inverted index in MapReduce
- Ranking: content score and its manipulation, page reputation.
- **HITS**: hubs and authorities, scores, iterative method, attacking HITS.
- **PageRank**: definition, flow formulation, power iteration, random walk interpretation, the Google formulation, *topic-aware PageRank*, link spam and attacking PageRank, *TrustRank*. 