Assignment 2 (Distributed Thursday March 16, 2017, Due: Tuesday April 25, 2017)

Submission Requirements

- All answers must be in pdf format.
- Please submit your assignment at Compass-2G
- Submission deadline: April 25, 2017

Question 1: Phrase Mining (25 points)

1. There are several phrase-based topic modeling methods, such as (1) Topical N-grams (TNG) (Wang, et al.’07), (2) TurboTopics (Blei & Lafferty’09), (3) KERT (Danilevsky, et al.’13), and (4) ToPMine (El-kishky, et al.’15). Compare and outline their major differences. What are the reasons that ToPMine may find quality phrases and phrase-based topics? (10 points)

2. SegPhrase+ (Liu et al, SIGMOD’15) does not use any natural language processing (NLP) methods in phrase mining. Give two examples and show NLP methods can be integrated into the Segphrase framework to further improve the quality of phrase mining. (15 points)

Question 2: Entity Recognition, Typing, Embedding and Network Construction (25 points)

1. Explain what challenges ClusType (Ren et al., KDD 2015) will encounter when being applied to solve fine-grained entity typing problem (Ren et al., KDD 2016) and explain why PLE can overcome these challenges when leveraging distant supervision. (10 points)

2. Discuss why jointly extract typed entities and relations may be helpful compared to solving each subtask independently. Explain how CoType (Ren et al., WWW 2017) is modeling entity types and relation types in a joint manner. Outline some other method designs that can also jointly model entity and relation types. (15 points)

Question 3: Truth Finding (25 points)

1. Explain what are the differences among the three truth finding algorithms:
TruthFinder (Xin et al., TKDE 2008), LTM (Zhao, et al, VLDB 2012) and GTM(Zhao, et al, QDB 2012). (10 points)

2. Not every piece of news or tweets is truthworthy. Design a mechanism that may use both sources to identify what is likely to be the truth in news and tweets. (15 points)

Question 4: Spatiotemporal and Social Media Data Mining (25 points)

1. Given the user movement database (time, location, location category), outline and describe a method to mine the semantic-rich movement patterns. (10 points)

2. Many tweets are geo-coded (i.e., their geo-locations are known). Suppose a tweet contains user-id, time, location, hashtag, and short text messages. Design an effective method that may detect an unusual local event from those that happen regularly in a local region. (15 points)