Efficient Data Parsing and Vandalism Detection on (Big) Knowledge Bases using Apache Spark and Hadoop ecosystem

Nayef Roqaya

**Motivation**
- Some people vandalize those knowledge bases.
- Patrollers are busy.
- Vandalism is not detected in time.

**Methods**
- **Distributed Data parser**
- **Distributed Vandalism Detector**

**Analysis**
- **Spark ML**
- ROC
- PR

**Results & Limitations**
- Features engineering plays a vital role in improving results of ML vandalism approach.
- By Spark with controlled partitioned techniques, the performance is efficient.
- Hadoop consumes a big size data storage.
- Streaming Data from Hadoop cannot deal with compressed files as result to need to keep **logic partitioning** and **physical partitioning** in our case study.

**Conclusion**
Vandalism can reduce the quality of knowledge bases. Detect vandalism automatically is possible.

**References:**