**Exercise**

**CSE5243 - Introduction to Data Mining**

**Jaccard Text Similarity**

**This is an in-class exercise for each group.**

* Be **respectful** to your group-mates.
* In a short exercise such as this one, you may not have time to answer the questions completely – just give it a quick try.
* **HAVE FUN!**

**Three Movies** (with sets of “Stars” listed):

* **Iron Man 1:** { Robert Downey Jr., Terrence Howard, Jeff Bridges, Shaun Toub, Gwyneth Paltrow }
* **Iron Man 2:** { Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Scarlett Johansson, Sam Rockwell, Mickey Rourke, Samuel L. Jackson }
* **Iron Man 3:** { Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, Rebecca Hall, Stéphanie Szostak, James Badge Dale, Jon Favreau, Ben Kingsley }

**Jaccard Index** (for two sets)**:**

**J (A, B) = |A intersect B| / |A union B|**

**Do the following:**

1. **Introduce** yourselves and get to know one another.
2. Choose a “speaker” who will present your findings to the rest of the class when the exercise is done. Specifically, choose the person whose **birthday is the furthest from the current date**.
3. Answer the following questions:
   1. What is the Jaccard Index for each pair of movies?
   2. Which pair of movies is most similar?
   3. Which movie earned the most at the box office? Should it have?
4. When the Breakout session ends, the “speaker” will be asked to present your findings. Please support them if needed.

If you are interested in this area, there are many online references. Here is one: <https://studymachinelearning.com/jaccard-similarity-text-similarity-metric-in-nlp/>