Dataset Explanation:

The dataset contains a real sample of 12 months logs (Mar'16 - Feb'17) from CI&T's Internal Communication platform (DeskDrop). It contains about 73k logged users interactions on more than 3k public articles shared in the platform. CI&T is a nearshore information technology and software engineering company based out of Brazil. It consists of two different parts :

**Shared\_articles.csv**

* This file contains information about the articles shared in the platform. Each article has its sharing date (timestamp), the original url, title, content in plain text, the article' lang (Portuguese: pt or English: en) and information about the user who shared the article (author).
* There are two possible event types at a given timestamp - (i) CONTENT SHARED: The article was shared in the platform and is available for users (ii) CONTENT REMOVED: The article was removed from the platform and not available for further recommendation.
* For the sake of simplicity, we only consider here the "CONTENT SHARED" event type, assuming (naively) that all articles were available during the whole one year period. For a more precise evaluation (and higher accuracy), only articles that were available at a given time should be recommended, but we let this exercice for you.

**Users\_interaction.csv**

This file contains logs of user interactions on shared articles. It can be joined to **articles\_shared.csv** by **contentId** column.

The eventType values are:

* **VIEW**: The user has opened the article.
* **LIKE**: The user has liked the article.
* **COMMENT CREATED**: The user created a comment in the article.
* **FOLLOW**: The user chose to be notified on any new comment in the article.
* **BOOKMARK**: The user has bookmarked the article for easy return in the future.

Probable tasks (Below pointers are for direction purpose only):

* **Data Munging**
* **Remove the cold start problem**
* **Smoothening user preference**
* **Create a class for modelling of a collaborative filtering recommender system**
* **Create an evaluation system using Top-N accuracy metrics through the following steps:**
* For each item the user has interacted in test set
* Sample 100 other items the user has never interacted.
* Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items
* Compute the Top-N accuracy metrics for this user and interacted item from the recommendations ranked list
* Aggregate the global Top-N accuracy metrics
* **Calculate the global metrics**

System design tasks (Below pointers are for direction purpose only):

• Design system architecture to deploy ML Model in production  
• How do you perform canary build?  
• What should be the strategy for ML Model Monitoring?  
• How do you perform load and stress testing?  
• How do you track, monitor and audit ML training?  
• Design framework for continuous delivery and automation of machine learning tasks

Deliverables:  
- Jupyter notebook (or equivalent) showcasing your work  
- Powerpoint presentation clearly explaining the approach and findings.  
- System design architecture (if applicable) and explanations.