# Entity Matching

Sowrabha Horatti Gopal  sgopal3@wisc.edu
Pradeep Kashyap Ramaswamy  pradeep.kashyap@wisc.edu
Sharath Prabhudeva Hiremath  shiremath@wisc.edu

### Items this Doc should cover

Provide a link to a pdf document that contains at least the following:

1. Describe the type of entity you want to match, briefly describe the two tables (e.g., where did you obtain these tables), list the number of tuples per table.
2. Describe the blocker that you use and list the number of tuple pairs in the candidate set obtained after the blocking step.
3. List the number of tuple pairs in the sample G that you have labeled.
4. For each of the six learning methods provided in Magellan (Decision Tree, Random Forest, SVM, Naive Bayes, Logistic Regression, Linear Regression), report the precision, recall, and F-1 that you obtain when you perform cross validation for the first time for these methods on I. Report which learning based matcher you selected after that cross validation.
5. Report all debugging iterations and cross validation iterations that you performed. For each debugging iteration, report
   (a) what is the matcher that you are trying to debug, and its precision/recall/F-1
   (b) what kind of problems you found, and what you did to fix them
   (c) the final precision/recall/F-1 that you reached.
6. For each cross validation iteration, report
   (a) what matchers were you trying to evaluate using the cross validation, and
   (b) precision/recall/F-1 of those.
7. Report the final best matcher that you selected, and its precision/recall/F-1.
8. It is important to note that all precision/recall/F-1 numbers asked for in the aboves are supposed to be numbers obtained via CV on the set I. Do not yet use set J.
9. Now report these numbers:
   For each of the six learning methods, train the matcher based on that method on I, then report its precision/recall/F-1 on J.
10. For the final best matcher Y selected, train it on I, then report its precision/recall/F-1 on J.
11. Report approximate time estimates:
    (a) to do the blocking
    (b) to label the data
    (c) to find the best matcher.
12. Provide a discussion on why you didn't reach higher recall, and what you can do in the future to obtain higher recall.
13. **BONUS POINTS:**
    (a) Provide comments on what is good with Magellan and what is bad, that is, as users, what else would you like to see in Magellan.
    (b) Are there any features/capabilities that you would really like to see being added? Any bugs?
    (c) Depending on how detailed and helpful these comments are, you can get bonus point from 1-10 (which will help with the final grade, not just with the project).

# 1 Entity

Books
Our entity is a tuple containing **Book Details**. We have scraped book details from the following two data sources -

1. **Goodreads**
2. **Barnes & Noble**

We chose these data sources as they give us comprehensive details of all the books.
Following table gives the schema of entities scraped from **Goodreads** and **Barnes & Noble**

| **Goodreads** | **Barnes & Noble** |
|---|---|
| Publication | Publisher |
| ISBN13 | ISBN-13 |
| Author | Author |
| Original_Title | Original_Title |
| Published_Date | Publication date |
| Pages | Pages |
| Ratings | – |
| Genres | – |
| Edition_Language | – |
| ISBN | – |
| Title | – |
| Reviews | – |
| Average_Rating | – |
| Edition | – |

Table 1: Entity Schema

- Number of tuples present in **Goodreads** table -
- Number of tuples present in **Barnes & Noble** table -

Final Schema being used -

| **Attributes** |
|---|
| Original_Title |
| Author |
| ISBN13 |
| Publication |
| Published_Date |

Table 2: Final Schema