Proposal Grading Policy

As Speakers:

- All members **get ready** when it comes to your turn.
- **Present for 5 minutes.**
- **Take just one question from students for 2 minutes.**

As Listeners:

- Listen to the presentation.
- Volunteer to **ask a question**: One question from all for one team.
- Take notes.
- **(After class) Write down your NetID. Give grades, questions, comments to each presentation.**
- **(On Feb 8 Thu) Return the handout to the instructor.**

As Instructor:

- Carefully read proposal papers and listen to the presentations.
- Carefully read grades, questions, comments given by students.
- Grade in scale of **{10, 9, 8, 7, 0}**. Members in a team will have the same grade. It takes 10% of the project and project takes 30% of the course – so, **10 = 3 points, 7 = 2.1 points of 100** in the final score.
    - Grades will be given on **Feb 8 Thu evening**.
    - *"For students who get 0, you will receive an email from me on Feb 8 evening. You will be asked to talk to me in my office. You can re-submit a proposal paper before Feb 14 Wed. You will be re-graded at {7, 6, 0}. If it's still 0, we need to talk more."*

Your NetID:

Your Name:

Proposal Presentation Schedule

Presentation & QA time	Team	Students			Title
Movie Session (2 projects)					
2:05-2:09, 2:10-2:11	NPM	jborrero	Borrero Cordova, Juan	Senior	The Netflix Problem: Movie Clustering and Classification Based on Ratings
		bhansen4	Hansen, Brandon	Senior	
		mprosser	Prosser, Mason	Senior	
2:12-2:16, 2:17-2:18	ACC	rmackey1	Mackey, Ryan	Senior	Actor Clustering and Cast Significance on Genre and Movie Ratings
		kshin1	Shin, Kevin	Senior	
Sports Session (4 projects)					
2:19-2:23, 2:24-2:25	MLB	abrizius	Brizius, Alex	Senior	Predicting MLB Performance Based on Minor League Statistics
		mburke18	Burke, Michael	Senior	
		momalle3	O'Malley, Michael	Senior	
		jdumford	Dumford, Jacob	Graduate-Masters	
2:26-2:30, 2:31-2:32	MML	sbanerj2	Banerjee, Sreya	Graduate-PhD.	Making March Less Mad - Predicting the NCAA Men's Basketball Tournament
		gwright3	Wright, Gabriel	Graduate-PhD.	
2:33-2:37, 2:38-2:39	EBM	nrao	Rao, Nathan	Junior	What Statistics are most Impactful? Examining Baseball's Metrics as Indicators for Success
		jspence5	Spencer, Joseph	Junior	
		rloizzo	Loizzo, Ryan	Junior	
		dchao	Chao, David	Junior	
2:40-2:44, 2:45-2:46	POW	salpteki	Alptekin, Samuel	Junior	Predicting the Outcome of Week 1 Collegiate Football Games
		jbeiter	Beiter, Jacob	Junior	
		sberning	Berning, Samuel	Junior	
		bshadid	Shadid, Benjamin	Junior	

Life Session (4 projects)					
2:47-2:51, 2:52-2:53	PBC	amital	Mital, Aman	Junior	Predicting Breast Cancer Diagnosis from Tumor Measurements
		anemecek	Nemecek, Andrew	Junior	
2:54-2:58, 2:59-3:00	DPH	wbadart	Badart, William	Senior	Determining predictors of H-1B salary and approval
		lduane	Duane, Luke	Senior	
		wyu1	Yu, Wenhao	Non-Degree Student	
3:01-3:05, 3:06-3:07	AFG	mgianni1	Giannini, Mark	Graduate-Masters	It's All Funds & Games - Predicting Kickstarter Success
		ptinsley	Tinsley, Patrick	Graduate-Masters	
		btunnell	Tunnell, Brian	Graduate-Masters	
3:08-3:12, 3:13-3:14	MPT	xwang41	Wang, Xueying	Graduate-PhD.	Misread-Proof Temporal Fact Extraction
		tzhao2	Zhao, Tong	Graduate-PhD.	

(from "Project instruction": http://www.meng-jiang.com/teaching/CSE647Spring18-Project.pdf)

The project proposal (proposal paper) will be graded as follows:

Title of Project:	**5%**	**What's the title of the project?**
Project Plan:	**30%**	**What do you plan to do?**
Data Sources:	**20%**	**What data do you plan to use? From where will this data come?**
Proposed Evaluation:	**30%**	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Writing Quality:	**15%**	**Clarity of expression (5%), organization (5%), and grammar (5%).**

2:05-2:09, 2:10-2:11	NPM	jborrero	Borrero Cordova, Juan	Senior	The Netflix Problem: Movie Clustering and Classification Based on Ratings
		bhansen4	Hansen, Brandon	Senior	
		mprosser	Prosser, Mason	Senior	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Team Members: Juan Borrero, Brandon Hansen, Mason Prosser

Project Proposal

The Netflix Problem: Movie Clustering and Classification Based on Ratings

Project Plan

The task is to determine what movie attributes correlate or impact its scoring. For this we plan to use unsupervised methods to help indicate which items might be relevant or irrelevant. In particular, this will be clustering, in which different movie attributes will be compared to their relative score in order to find group correlations between them. This will function as a form of exploratory data analysis that will help us build metrics to determine a movies likely score without knowing its score beforehand. The clustering methods used in particular will be determined based on the project's practical necessity and it appropriateness to the task.

Data Sources

The dataset we plan to use is a Kaggle dataset and its name is: The Movies Dataset. The dataset is a CSV file and has 228MB worth of data. It consists of movies released on or before July 2017 and contains metadata on over 45 thousand movies with more than 26 million reviews from over 270 thousand users. The metrics and data points in the dataset are: ratings from 1 to 5, cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. A link to the Dataset is found below:

- https://www.kaggle.com/rounakbanik/the-movies-dataset

Evaluation

We will evaluate the success of our methods by observing the clusters that appear and testing to see if dummy movies inserted into the data are sorted into the expected clusters. Classification may be evaluated by dividing the data into and testing sets and comparing predicted ratings to the actual ratings.

2:12-2:16, 2:17-2:18	ACC	rmackey1	Mackey, Ryan	Senior	Actor Clustering and Cast Significance on Genre and Movie Ratings
		kshin1	Shin, Kevin	Senior	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Project Proposal: Actor Clustering and Cast Significance on Genre and Movie Ratings

Ryan Mackey and Kevin Shin

1 INTRODUCTION

Given a data set regarding movies (rating, director, actors), can the genre and success of a movie be determined? And can meaningful clusters of people within the industry be determined?

1. Group actor/actress/directors into a clustering to see similarity in genre.
2. Categorize a movie into a genre based on who the director is and who is casted.
3. Determine success of movie based on director and cast and compare to the rating of the movie.

The methodology for determining each of these steps will go as follows:

1. Clustering: The k-medoids algorithm [1] will be used to determine groups of actors who work closely together. We will use the silhouette tactic to assist in determining the value of *k* before running the algorithm. We will retroactively examine select clusters to assign useful titles to the groupings.
2. SVM/Decision Tree: We will try boh methods to determine which algorithm is better at classifying genres and success of movies. 80-90% of our data set will be used as training data and the remaining will be used as testing data. For genre determination, the model will be trained by cast names, director, and genre. For movie rating predictions, the same data will be used but instead of the genre, it will be the movie rating. For both classifications, inputs for testing will be the names of the actors and directors.

2 RELATED WORKS

3 PROBLEM DEFINITION

4 PROPOSED METHODOLOGY

Based on our SVM and/or Decision Tree, we can determine the accuracy of genre predictions and movie rating predictions. With genre, we will observe how many times the model detects the correct genre for the movies and output an accuracy percentage. With movie ratings, we will collect the variance of the outputted predicted result and use that to determine how precise our model is. Based on the accuracy and sample size, we can see whether cast and crew

can be used in determining the genre or the average rating of a movie.

For the clustering problem, distances between people in the industry will be determined based on their association (movies worked on together, mutual connections, etc.) and the k-medoids algorithm will divide them into clusters as appropriate. We will try several k values, but we will inform ourselves initially with the silhouette method. Once we are satisfied with the outcome, we will examine a selection of the clusters and consider why they are clustered together (genre, age, relationships etc.).

5 DATA AND EXPERIMENTS

5.1 Data Set

Our data will be drawn from IMDb (http://www.imdb.com/interfaces/). The information is broken up into 7 databases containing different information about the movies or actors in each. The movie unique identifier serves as the primary key for 6 of the databases, while the actor/actress/director unique identifiers serve as a foreign key in those 6 databases, and as the primary key in the 7th. Relevant data includes movie titles, top billed cast and crew on each, average IMDb ratings, and genre information.

5.2 Experimental Settings

5.3 Evaluation Results

6 CONCLUSIONS

REFERENCES

[1] Hae-Sang Park, Chi-Hyuck Jun. 2008. *A simple and fast algorithm for K-medoids clustering*

2:19-2:23, 2:24-2:25	MLB	abrizius	Brizius, Alex	Senior	Predicting MLB Performance Based on Minor League Statistics
		mburke18	Burke, Michael	Senior	
		momalle3	O'Malley, Michael	Senior	
		jdumford	Dumford, Jacob	Graduate-Masters	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Data Science Project Proposal
Submitted: February 5th, 2018

Predicting MLB Performance Based on Minor League Statistics

By:
Alex Brizius - Michael Burke - Jacob Dumford - Michael O'Malley

Plan:

For this project, we will be using Minor League baseball statistics to predict the "success" of a player in their MLB career. Baseball is particularly suited to this sort of analysis for a variety of reasons:

- Data for both Major League players and Minor League players is readily available (see **Data Sources** for an enumeration of sources that will be used).
- Baseball has a rich history of data collection / analyzation that will facilitate this project.
- Likewise, baseball has an extremely developed Minor League system that provides young prospects the opportunity to develop while playing against competition of a similar level. This is in contrast to other major American sports, where prospects go directly to the highest level of play.

Because of the massive amount of data points collected per player, it will be necessary for the experimenters to limit the scope of the project. Namely, we will analyze exclusively the offensive (batting and baserunning) statistics of outfield (non-pitching) players. Further, we plan to only consider statistics recorded at an A through AAA level of the MiLB farm system. Although there are more professional and semi-professional teams that do not fall into this category, a large portion of MLB players will have spent the majority of their development at this level. Likewise, we will not analyze college or international statistics to help limit our dataset.

Data Sources:

There are a multitude of existing repositories to find both MLB and MiLB dat, including:

- baseball-reference.com
- mlb.com
- milb.com
- sabr.org

In general, we anticipate that for this project that cleaning, integrating, and selecting proper data points will be much more difficult than simply finding the raw data.

Method Evaluation:

A large portion of this project will rely on proper pruning of the vast amount of raw data available. For example, it will be necessary to cull the number of players under consideration by numerous metrics, such as years active and minimum plate appearances. Further, there are some data points that are available for MLB players but likely not MiLB, such as exit velocity on batted balls, batting order, and stadium data. As such, although such statistics may be an effective means of evaluating player performance, it cannot be relied upon for our analysis. Likewise, gauging MLB "success" is a somewhat fuzzy objective: should subjective assessments such as Hall of Fame and All-Star voting be considered, or should objective measurements exclusively be used? This question will be answered as we begin our analysis.

In addition, there are complicating factors that may be ignored to simplify the problem. For example, teams general value batting statistics differently based on the outfield position of the player. In this way, evaluating the offensive capabilities of a catcher (the most skilled defensive position) is very different than that of a right fielder (where players who are offensively skilled but defensively limited are often placed). Similarly, left-handed batters (almost) exclusively play first base and the outfield, because the remaining defensive positions are better suited for right-handed throwers. As such, there are endless intricacies that our method of evaluation could attempt to embed, but we will clearly need to carefully choose what we consider in order to keep the scope of the project feasible.

Further, correlation analysis will be critical for limiting the size of our data set. There are numerous statististics in baseball that overlap in what they measure. Practically, it may not be necessary to consider each of batting average, on-base percentage, and slugging percentage for our analysis for example. In this way, more advanced metrics such as WAR (Wins Above Replacement) and Wins Shares may prove extremely useful, as they can help simplify numerous features into one summary feature.

To determine whether our analysis is successful, our dataset will be split into a training portion and a testing portion; after developing the evaluation on the training subset, it will be applied to a testing portion of MiLB player data to predict the success of those players. Then, the prediction of each player's "success" in the MLB can be compared to their actual career statistics to determine to what extent the method was successful.

2:26-2:30, 2:31-2:32	MML	sbanerj2	Banerjee, Sreya	Graduate-PhD.	Making March Less Mad - Predicting the NCAA Men's Basketball Tournament
		gwright3	Wright, Gabriel	Graduate-PhD.	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Making March Less Mad - Predicting the NCAA Men's Basketball Tournament

Sreya Banerjee
University of Notre Dame
Notre Dame, Indiana
Sreya.Banerjee.9@nd.edu

Gabriel Wright
University of Notre Dame
Notre Dame, Indiana
Gabriel.S.Wright.142@nd.edu

Abstract

March madness is around the corner where the top 64 college basketball teams from across America participate to win the national championship. For basketball enthusiasts and computer scientists alike, in essence, this becomes a binary classification task where, with sufficient historical data, we can try to predict the outcome of tournament. In this work, we plan to implement and evaluate common supervised machine learning approaches to predict the result of the 63 games within the tournament, based on a many years worth of basketball results. The project would involve common data science tasks like data integration, augmentation, cleaning, feature extraction, feature selection, model building and tuning. Some algorithms we wish to test are: logistic regression, decision tree, Naive Bayes and support vector machine. For evaluation we plan to use the classification accuracy as a metric.

1. Introduction

Given two data sets: one being a data set of the outcome of NCAA college basketball games from previous years, and the other containing individual statistics for each team, is it possible for machine learning algorithms to predict the results of the 63 games in the tournament? Essentially, the task can be formulated as a binary classification problem with two possible outcomes of win or lose for each game. A "successful" model would be one that outperforms the null model where each game is picked based on tournament seeding.

For the classification task, we plan to use the common machine learning algorithms like Naive Bayes, support vector machines [1] and decision trees. We chose the data set [1] where the results of basketball matches each year is tabulated as csv file with the attributes described already. Additionally, the dataset [2] will provide advanced statistics for each team in each year. We plan to use a decade worth of data. However, since the results of 2016, 2017, 2018 are not present, we plan to create them ourselves. The data from 2018 would be solely used for prediction purpose.

For the tasks, we will divide the data set into training (all the historical data) and testing (the 2018 tournament) components. We will use the training set to fine-tune the model through cross-validation evaluation on the testing set. To select the best classification model, we plan to use the classification accuracy metric on the validation data set. Additionally, because the team data has many features features, we would look into some feature extraction and selection strategies like principal component analysis. For decision trees, we plan to evaluate which features among the 12 described above are actually relevant for classification tasks through ID_3 (information gain), $C_{4.5}$ (gain ratio) and CART(gini index).

[1] https://github.com/octonion/basketball/tree/master/ncaa/csv

[2] https://www.basketball-reference.com/

2. Related Work

Review classification papers/code packages/tools from [2]

3. Proposed Methodology

4. Data and Experiments

4.1. Data set

4.2. Experimental settings

4.3. Evaluation results

5. Conclusion

References

[1] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[2] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

2:33-2:37, 2:38-2:39	EBM	nrao	Rao, Nathan	Junior	What Statistics are most Impactful? Examining Baseball's Metrics as Indicators for Success
		jspence5	Spencer, Joseph	Junior	
		rloizzo	Loizzo, Ryan	Junior	
		dchao	Chao, David	Junior	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Project Proposal: What Statistics are most Impactful?

Examining Baseball's Metrics as Indicators for Success

Nathan Rao, Joseph Spencer, Ryan Loizzo, DJ Chao

ACM Reference format:
Nathan Rao, Joseph Spencer, Ryan Loizzo, DJ Chao. 2018. **Project Proposal: What Statistics are most Impactful?**. In *Proceedings of Data Science 40647, Notre Dame, IN USA, Spring 2018,* 1 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

Given a comprehensive database of major league baseball's raw and advanced metric statistics, how accurately can specific team and individual statistics determine the outcome for any given team's season?

The data will primarily be directed at determining the final outcome of the season - the World Series champion. However, aside from the main task of determining the champion, the data may be directed to the postseason as a whole, examining the statistical match-up to potentially find the "Cinderella Story" spoken about so often across all of the major sports.

We will pull complete team and individual statistics for analysis. The variety of data will give us the greatest accuracy in determining which statistics have the greatest meaning come season's end. This data will be primarily pulled from Baseball-Reference (https://www.baseball-reference.com/), one of the largest baseball statistic compilation websites on the internet.

We will evaluate our final method by applying our final algorithm to previous seasons and seeing how accurately it predicts the end of the regular season standings and postseason outcomes. Correctly predicting the teams that move on to the postseason and win the championship will deem our method and algorithm successful.

Finally, determining if our method is successful is relatively straightforward. If the predictions are accurate, as in predicting the correct teams that move on to the postseason and win the championship, our method and algorithm will be successful.

Finally, with the successful completion of our analysis on baseball statistics and their ability to predict season outcomes throughout the league, this same method can be applied to determine a host of other sports related outcome. Football, basketball, hockey, and others all utilize advanced metrics, and this data could potentially be used in the same way as the compiled MLB data. Additionally, the individual statistics may help in determining special individually directed outcomes, such as the Hall of Fame. All of these other analyses could be the subject of future work from our original analysis.

Data Science 40647, Spring 2018, Notre Dame, IN USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

2:40-2:44, 2:45-2:46	POW	salpteki	Alptekin, Samuel	Junior	Predicting the Outcome of Week 1 Collegiate Football Games
		jbeiter	Beiter, Jacob	Junior	
		sberning	Berning, Samuel	Junior	
		bshadid	Shadid, Benjamin	Junior	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Predicting the Outcome of Week 1 Collegiate Football Games

Alptekin, Sam; Beiter, Jacob; Berning, Sam; Shadid, Ben
{salpteki,jbeiter,sberning,bshadid}@nd.edu

1 INTRODUCTION

Given team performance and recruiting data from the previous year, can a machine accurately predict the outcome of a Week 1 college football game?

This task is a classification problem — given a set of labeled data organized by game, build a model that can decide which team will win by comparing different statistical categories.

To solve this problem, we will be using multiple data sets such as S&P+ rankings [1], 247sports.com recruiting rankings [2], ESPN game scores from Week 1 of previous seasons [3], and NCAA statistics [4]. Data exists from each of these sources from 2013-present. We will need to compile these into our own database using web scraping and data integration techniques.

We plan to try several different models including Decision Trees and Support Vector Machines, since this task involves only classification with labeled data.

In addressing the classification problem, we will divide our data sets into training and testing portions. The training set will be used to learn classification models and perform rotation estimation evaluation on the testing set. By comparing our predicted outcomes to the real world outcomes of games in our testing set, we will be able to determine the accuracy of our model.

REFERENCES

[1]http://www.footballoutsiders.com/stats/ncaa
[2]https://247sports.com/Season/2018-Football/CompositeTeamRankings
[3]http://insider.espn.com/college-football/schedule
[4]http://stats.ncaa.org/rankings/change_sport_year_div

2:47-2:51, 2:52-2:53	PBC	amital	Mital, Aman	Junior	Predicting Breast Cancer Diagnosis from Tumor Measurements
		anemecek	Nemecek, Andrew	Junior	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Predicting Breast Cancer Diagnosis from Tumor Measurements

I. Introduction

Using data on Wisconsin breast cancer diagnosis, we plan to create a model which will classify a tumor as malignant or benign based on a series of measurements. Given the set of measurements, the model should output the class of the tumor.

We will be using a dataset provided by Kaggle from UC Irvine Machine Learning (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data).

To create the model, we will divide the dataset of 570 patients into two pieces, one for training and one for evaluation. The dataset contains 30 features in addition to the class and a patient identifier, so we will first attempt to reduce the dimensionality of the data to simplify the model.

2:54-2:58, 2:59-3:00	DPH	wbadart	Badart, William	Senior	Determining predictors of H-1B salary and approval
		lduane	Duane, Luke	Senior	
		wyu1	Yu, Wenhao	Non-Degree Student	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Determining predictors of H-1B salary and approval

Luke Duane, Wenhao Yu, & Will Badart

I. Introduction

An H-1B visa is a visa given out to foreign workers filling a specialty occupation for an American company. They are three years at minimum, but extendable to up to six years. Just last year in 2017, almost 350,000 foreign workers applied to come over and a little under 200,000 were approved.

The H-1B lottery is a laborious and complex process for both large companies bringing in thousands of migrant employees and small ones only onboarding a couple. A tool which highlights the important features that support H-1B approvals could be a vital strategic asset for these companies. Lots of data exists in this domain, but to integrate it and perform meaningful analysis is beyond the capabilities of companies without established data science practices.

We plan to produce a model that shows what features are most valuable in regards to H-1B workers' salaries and approval. First, we intend to implement a decision tree to focus on visa approvals and then also develop a neural network that utilizes all the features from the dataset to predict salaries.

II. Related Work

In April of 2017, Glassdoor published an article analyzing the salaries of H-1B immigrants and comparing them to those of domestic workers in similar roles and fields. While the report does not attempt to model H-1B workers' salaries based on other features, it offers a comprehensive statistical analysis of their pay. See: Glassdoor Comparison on H-1B Visa Salaries vs US Workers.

III. Problem Definition

Section A: H-1B Visa approvals and wage level

1. What are the most popular occupations for foreign employees? Given the occupation and expected income, how to predict the success rate of getting a H1-B Visa? (classification)

2. What is the average wage level of foreign employees? Given the occupation, can we predict the possible salary range? (classification)

Section B: Data Science related job position Analysis

1. Can we analyze the number of jobs related to data analysis from complex occupational categories? (classification)

2. According to the time series model, how can we predict the number of data scientists and wage trends? As the number of H1-B visas is limited every year, can we analyze the possibility that data scientists will be more likely to get jobs and H1-B visa and have higher income than the past? (prediction)

Section C: Location-based analysis for different jobs

1. How can we get occupations distribution map in the U.S? Given the occupation, can we analyze where(states or city) can they easier to get a related job? (clustering)

2. Can we use the relevant algorithms to give more precise information about the work?

IV. Proposed Methodology

1. Data Cleaning: Missing and noisy data cleaning

   There are many missing data and noisy data in the dataset, so we need to some data cleaning work before we analyze them. There are some examples as following:

	A	B	C	D	E	F	G	H	I	J	K
1		CASE_STATU	EMPLOYER_	SOC_NAME	JOB_TITLE	FULL_TIME_	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
2	7501	DENIED	LOTUS SOFT	COMPUTER	CIS MANAG	N	0	2016	NEW YORK, NEW YORK	-74.005941	40.7127837
3	22308	DENIED	PERSON AR	MANAGERS	E - COMMEI	N	0	2016	PISCATAWAY, NEW JERSE	-74.464286	40.554887
4	52504	DENIED	NATIONAL	BUSINESS O	OPERATION	N	0	2016	CHICAGO, ILLINOIS	-87.629798	41.8781136
5	57486	DENIED	SARAS FAM	ACCOUNTA	ACCOUNTIN	N	0	2016	FORT WAYNE, INDIANA	-85.139351	41.079273
6	193691	DENIED	NESS USA, I	COMPUTER	PROGRAMN	N	0	2016	PITTSBURGH, PENNSYLVA	-79.995886	40.4406248
7	223688	WITHDRAW	TECHMATRI	COMPUTER	COMPUTER	N	0	2016	NORTH BRUNSWICK, NEW	-74.476671	40.4525163
8	535858	DENIED	EOS UNIFIEI	COMPUTER	FIELD ENGI	N	0	2016	SANTA CLARA, CALIFORN	-121.95524	37.3541079
9	569246	DENIED	SEC SOLUTI	ARCHITECTU	ARCHITECTU	N	0	2016	LIBERTY HILL, TEXAS	NA	NA
10	595705	DENIED	MCNAMARA	LAWYERS	ATTORNEY	N	0	2016	NAPLES, FLORIDA	-81.79481	26.1420358
11	611583	DENIED	SUPREME C	LIBRARIANS	LAW LIBRAI	N	0	2016	NEW YORK CITY, NEW YO	-74.005941	40.7127837
12	620655	DENIED	APOLLO ELI	PUBLIC REL	CORPORATI	N	0	2016	NEW YORK, NEW YORK	-74.005941	40.7127837
13	643551	DENIED	T & F MAIN	CHEFS AND	CHEF - PRO	N	0	2016	SPARTANBURG, SOUTH C	-81.932048	34.9495672
14	644650	DENIED	LEGACY INT	REAL ESTAT	FOREIGN RE	N	0	2016	MIAMI, FLORIDA	-80.19179	25.7616798
15	677394	DENIED	HARRIS THE	SPEECH-LA	SPEECH LAN	Y	0	2015	HONOLULU, HAWAII	-157.85833	21.3069444
16	739878	DENIED	STOBI, LLC	ENGINEERS,	COMMISSIO	Y	0	2015	TAMPA, FLORIDA	-82.457178	27.950575
17	885748	DENIED	WESTERN P	MECHANICA	MECHANICA	Y	0	2015	EUGENE, OREGON	-123.08675	44.0520691
18	953671	DENIED	JONATHAN	DENTISTS, G	DENTIST	Y	0	2015	GARLAND, TEXAS	-96.638883	32.912624
19	612562	DENIED	CENTRO CU	ART DIRECT	CULTURAL I	N	35	2016	MIAMI, FLORIDA	-80.19179	25.7616798

Missing data: many places of salary are 0

	A	B	C	D	E	F	G	H	I	J	K
1048541	631985	CERTIFIED	WESTERN K	PHYSICIANS	HEMATOLO	Y	255154	2016	PADUCAH, KENTUCKY	NA	NA
1048542	626324	CERTIFIED	APOGEE ME	INTERNISTS	HOSPITALIS	Y	256526.4	2016	PORTAGE, WISCONSIN	NA	NA
1048543	628150	CERTIFIED	MERCY PHY	INTERNISTS	CARDIOLOG	Y	272750	2016	EUNICE, LOUISIANA	NA	NA
1048544	775757	DENIED	SOCCER AN	ENTERTAINI	SOCCER TRA	Y	276000	2015	ATLANTIC BEACH, NEW Y	NA	NA
1048545	628033	CERTIFIED	DUBOIS REG	INTERNISTS	PULMONOL	Y	276016	2016	PHILIPSBURG, PENNSYLV	NA	NA
1048546	631792	CERTIFIED	PHYSICIANS	PHYSICIANS	ONCOLOGI	Y	278733	2016	NEW ULM, MINNESOTA	NA	NA
1048547	632981	CERTIFIED	ALLINA HEA	PHYSICIANS	MEDICAL OI	Y	292138	2016	FARIBAULT, MINNESOTA	NA	NA
1048548	745530	CERTIFIED-V	LAUREL EYE	SURGEONS	VITREORETII	Y	307211	2015	CLEARFIELD, PENNSYLVA	NA	NA
1048549	745660	CERTIFIED-V	LAUREL EYE	SURGEONS	VITREORETII	Y	307211	2015	BROOKVILLE, PENNSYLVA	NA	NA
1048550	786821	CERTIFIED	LAUREL EYE	SURGEONS	VITREORETII	Y	307211	2015	CLEARFIELD, PENNSYLVA	NA	NA
1048551	786940	CERTIFIED	LAUREL EYE	SURGEONS	VITREORETII	Y	307211	2015	BROOKVILLE, PENNSYLVA	NA	NA
1048552	630383	CERTIFIED	REGIONAL I	PHYSICIANS	GENERAL SU	Y	315668	2016	KANE, PENNSYLVANIA	NA	NA
1048553	788666	CERTIFIED	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	POPLARVILLE, MISSISSIPP	NA	NA
1048554	789078	CERTIFIED	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	LAUREL, MISSISSIPPI	NA	NA
1048555	792858	CERTIFIED-V	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	POPLARVILLE, MISSISSIPP	NA	NA
1048556	793146	CERTIFIED-V	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	LAUREL, MISSISSIPPI	NA	NA
1048557	827259	CERTIFIED	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	POPLARVILLE, MISSISSIPP	NA	NA
1048558	829414	CERTIFIED	HATTIESBUF	PHYSICIANS	VASCULAR S	Y	324694	2015	LAUREL, MISSISSIPPI	NA	NA
1048559	699385	CERTIFIED	HATTIESBUF	PHYSICIANS	CARDIOLOG	Y	325119	2015	PRENTISS, MISSISSIPPI	NA	NA

Missing data: many places of location are NA

2. Data cleaning: Extreme data cleaning

   When we analyze the average income level, we need to remove some extreme values; otherwise we will have a deviation in the wage prediction.

3. Classification: job title

   (1) Description: In the original dataset, we can see that there are a lot of job titles, but people tend to focus on general classification, such as data scientists, company manager, etc.

   We will firstly find the key words corresponding to data scientists (also analyzing other occupations). Ex. Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. Next, we will use these key words to classify the job titles into general name.

   (2) Method: In pattern recognition, the k-nearest neighbors algorithm(k-NN) is a non-parametric method used for classification and regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

A	B	C	D	E	F	G	H	I	J	K
43128	CERTIFIED	HAKKASAN	MEETING, CONVEN	EVENTS MANAGER	N	35838	2016	HONOLULU, HAWAII	-157.85833	21.3069444
45201	CERTIFIED	JTB GLOBAL	MARKET RESEARCI	KOREAN MARKET RESEARCH ANALYST	N	36213	2016	HONOLULU, HAWAII	-157.85833	21.3069444
45446	CERTIFIED	PRTECH LLC	MARKET RESEARCI	ONLINE MARKETING & CONTENT DEV'T SPECIA	N	36213	2016	HONOLULU, HAWAII	-157.85833	21.3069444
843844	CERTIFIED	WINCUBIC.(	MARKET RESEARCI	MARKET RESEARCHER	N	36732.8	2015	HONOLULU, HAWAII	-157.85833	21.3069444
725764	CERTIFIED	WAIKIKI RES	MARKET RESEARCI	ASIA MARKET RESEARCH ANALYST	Y	36733	2015	HONOLULU, HAWAII	-157.85833	21.3069444
749653	CERTIFIED	PACIFIC TRA	MARKET RESEARCI	MARKETING ANALYST	Y	36733	2015	HONOLULU, HAWAII	-157.85833	21.3069444
835037	CERTIFIED	PRTECH LLC	MARKET RESEARCI	ONLINE MARKETING SPECIALIST - JAPANESE	Y	36733	2015	HONOLULU, HAWAII	-157.85833	21.3069444
25277	CERTIFIED	GOOD LUCK	HUMAN RESOURC	HR SPECIALIST	N	37003	2016	HONOLULU, HAWAII	-157.85833	21.3069444
697291	CERTIFIED	RESEARCH (	BIOCHEMISTS AND	POSTDOCTORAL RESEARCH FELLOW	Y	37793.6	2015	HONOLULU, HAWAII	-157.85833	21.3069444
54523	CERTIFIED	ALAKA'I VE	ACCOUNTANTS AI	ACCOUNTANT	N	38002	2016	HONOLULU, HAWAII	-157.85833	21.3069444
55360	CERTIFIED	ECA LLP FK/	ACCOUNTANTS AI	STAFF ASSOCIATE/ACCOUNTANT	N	38002	2016	HONOLULU, HAWAII	-157.85833	21.3069444
54201	CERTIFIED-\	UNIVERSITY	ACCOUNTANTS AI	INSTITUTIONAL SUPPORT	N	38148	2016	HONOLULU, HAWAII	-157.85833	21.3069444
624145	CERTIFIED	UNIVERSITY	DIETITIANS AND N	RESEARCH SUPPORT	N	38148	2016	HONOLULU, HAWAII	-157.85833	21.3069444
662190	CERTIFIED-\	UNIVERSITY	COMPUTER USER S	INFORMATION TECHNOLOGY SPECIALIST	Y	38148	2015	HONOLULU, HAWAII	-157.85833	21.3069444
732226	CERTIFIED-\	UNIVERSITY	RESIDENTIAL ADVI	INSTRUCTIONAL & STUDENT SUPP (RESIDENCE	Y	38148	2015	HONOLULU, HAWAII	-157.85833	21.3069444
575099	CERTIFIED-\	RESEARCH (	ZOOLOGISTS AND	JIMAR PIFSC ECOSYSTEMS RESEARCHER	N	38542.4	2016	HONOLULU, HAWAII	-157.85833	21.3069444
663771	CERTIFIED-\	RESEARCH (	ZOOLOGISTS AND	PMNM RESEARCH SPECIALIST	Y	38542.4	2015	HONOLULU, HAWAII	-157.85833	21.3069444
957212	CERTIFIED	ELDERS INTI	FOOD SCIENTISTS	FOOD SCIENTIST	N	38625.6	2015	HONOLULU, HAWAII	-157.85833	21.3069444
967571	CERTIFIED	FLYING FOC	FOOD SCIENTISTS	FOOD SAFETY & QUALITY TECHNICIAN	Y	38626	2015	HONOLULU, HAWAII	-157.85833	21.3069444
952391	CERTIFIED	SAN LOREN	MULTIMEDIA ARTI	MEDIA ARTIST:JAPANESE MARKET	N	38708.8	2015	HONOLULU, HAWAII	-157.85833	21.3069444
659799	CERTIFIED-\	RESEARCH (	CHEMISTS	LABORATORY/RESEARCH DATA TECHNICIAN	Y	38771.2	2015	HONOLULU, HAWAII	-157.85833	21.3069444
667348	CERTIFIED	GOOD LUCK	PUBLIC RELATION	PUBLIC RELATIONS SPECIALIST	Y	38792	2015	HONOLULU, HAWAII	-157.85833	21.3069444

Original job titles in the dataset

general name	Company manager	Data Scientist
Job title 1	CEO	PROGRAMER ANALYST
Job title 2	CHIEF EXECUTIVE OFFICER	SYSTEMS ENGINEER
Job title 3	CHIEF BUSINESS OFFICER	COMPUTER ANALYST
Job title 4	EXECUTIVE DIRECTOR	IT SPECIALIST
Job title 5	CHIEF OPERATING OFFICER	COMPUTER CONFIGURER
Job title 6	PRESIDENT, LENDING SOLUTIONS	PROGRAMMER ANALYST - II
Job title 7	PRINCIPAL	DATA ANALYST

Classified job titles

4. Prediction: salary level

   (1) Description: If given the occupation and location to predict the possible salary range, it is impossible to give an exact number. Because the value of salary is extremely variable, we can only divide the salaries into some intervals an give a range of salary level.

   We will build the salary decision tree. When given the occupation and location, it can predict the salary level.

   (2) Method: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

5. Prediction: data Scientist Job

   (1) Description: The dataset of H1-B Visa is from 2011 to 2016, so we can predict the future development of a specific occupation(Data Scientist) based on a time-series model, to

find the variables and test their stationary characteristics. Eventually, we will establish a regression model. It may be a linear model, an exponential model or an ARIMA model…

(2) Method: Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

6. Clustering: geographic distribution

(1) Description: We will associate geographic locations with jobs and use association rule learning to make a clustering. For example, we will make an connection between people who work in Wall Street and their jobs.

(2) Method: Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

V. Data Sets

1. One of the largest freely available datasets on H-1B applications comes from kaggle.com. It contains over 3 million records and tracks 10 different features per application. See: https://www.kaggle.com/asavla/h1-visa/data. This data covers applications roughly between 2012 and 2016.

2. Another key dataset comes from the Foreign Labor Certification Data Center. Its data is organized by year, spanning from 2001 to 2007. See: http://www.flcdatacenter.com/CaseH1B.aspx.

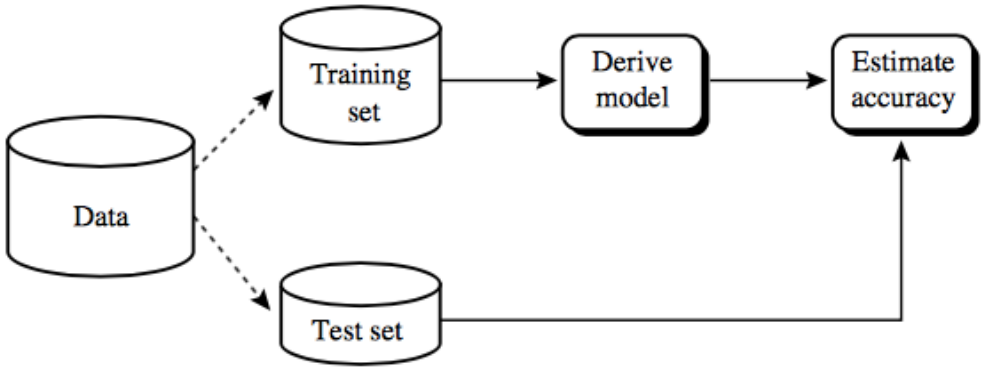
3. Besides, OFLC's annual reports also provides a lot of program information and data. Although it is not raw data, it disclosures cumulative quarterly and annual releases of program to assist with external research and program evaluation.

https://www.foreignlaborcert.doleta.gov/pdf/OFLC_Annual_Report_FY2016.pdf

VI. Evaluation Method and Results

1. The holdout method: We will randomly divide the dataset into two independent sets, a training set and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set.

In total, there are 3 million data in the dataset. So we have enough data to partition off a good-sized test set, which we will use to evaluate the accuracy of our predictive models.



2. Random subsampling: It is a variation of the holdout method in which the hold- out method is repeated k times.

We will use the overall accuracy estimate as the average of the accuracies obtained from each iteration.

3. Cross-validation: Leave-one-out is a special case of k-fold cross-validation where k is set to the number of initial tuples. That is, only one sample is “left out” at a time for the test set. In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.

We will stratify 10-fold cross-validation to estimate accuracy due to its relatively low bias and variance.

4. Using Statistical Tests of Significance such as ***t***-test.

VII. Conclusions

VIII. References

[1] The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011-3

3:01-3:05, 3:06-3:07	AFG	mgianni1	Giannini, Mark	Graduate-Masters	It’s All Funds & Games - Predicting Kickstarter Success
		ptinsley	Tinsley, Patrick	Graduate-Masters	
		btunnell	Tunnell, Brian	Graduate-Masters	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

It's All Funds & Games

Predicting Kickstarter Success

Mark Giannini
University of Notre Dame
mgianni1@nd.edu

Patrick Tinsley
University of Notre Dame
ptinsley@nd.edu

Brian Tunnell
University of Notre Dame
btunnell@nd.edu

1. Introduction

1.1 Project Plan

For our semester project, we have decided to test our ability to predict whether or not a given Kickstarter campaign will be successful. In order to be deemed a success, the proposed campaign needs to meet or exceed the funding goal proposed by the initial author by a predefined deadline; anyone can contribute as long as the campaign is still active. In the context of our project, each instance in our data set has a unique project ID and fourteen features; these include project name, project description, keywords, financial goal in US dollars, the project deadline and the number of backers contributing to and supporting the project. Using sentiment analysis and other logistic regression techniques we have learned in previous classes, we plan to predict the binary final_status field, which indicates a successful project (1) or a failed attempt (0).

1.2 Data Sources

Initially, we planned to crawl the data from the Kickstarter website ourselves. However, upon browsing a plethora of Kaggle competitions, we found a pre-built data set that contains all our fields of interest. The supplied data has 108,129 rows, each corresponding to a project proposal submitted between May 2009 and May 2015. Each instance has the following features: Project ID, Name, Description, Funding Goal, Project Keywords, Disable Communication, Country, Currency, Deadline Date, Date Created, Date Launched, State Changed At, Launched At, Number of Backers and finally, the targeted response variable, Final Funding Status.

1.3 Proposed Evaluation

To evaluate our models predictive power, we plan on splitting our data into two sets. The first partition will be the training set, and it will be used to build and train our model. The second partition will be the testing set, and it will be used to validate the model. If we split the data 70%-30% respectively, the training set will have 75,690 rows, and the testing set will have 32,439 rows. By withholding a subset of the full data set, we have the power to test our final model on unseen data, which can be used to evaluate estimator performance; this technique also helps to avoid overfitting, producing a more generalized model capable of predicting the success of future Kickstarter projects.

CONF 'yy, Month d–d, 20yy, City, ST, Country.

http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

2. Related Work

3. Problem Definition

4. Proposed Methodology

5. Data & Experiments

5.1 Data Set

5.2 Experimental Settings

5.3 Evaluation Results

6. Conclusions

A. Appendix Title

Appendix, if needed.

Acknowledgments

Acknowledgments, if needed.

3:08-3:12, 3:13-3:14	MPT	xwang41	Wang, Xueying	Graduate-PhD.	Misread-Proof Temporal Fact Extraction
		tzhao2	Zhao, Tong	Graduate-PhD.	

Title of Project:	**What's the title of the project?**
Grade (1-10):	
Project Plan:	**What do you plan to do?**
Grade (1-10):	
Data Sources:	**What data do you plan to use? From where will this data come?**
Grade (1-10):	
Proposed Evaluation:	**How do you plan to evaluate your proposed method? How will you determine whether the method is successful?**
Grade (1-10):	
Writing Quality:	**Clarity of expression, organization, and grammar.**
Grade (1-10):	

Other questions/comments:

Misread-Proof Temporal Fact Extraction

Xueying Wang, Tong Zhao, Meng Jiang
University of Notre Dame, Notre Dame, Indiana, 46556, USA
{xwang41, tzhao2, mjiang2}@nd.edu

ABSTRACT

...

1 INTRODUCTION

Thanks to high accuracy of entity typing systems [2, 4], typed textual pattern-based methods have been successful in extracting (entity, attribute, value)-tuples (called *EAV-tuples* or *facts*) from unstructured datasets such as news, tweets, and scientific publications [1, 3]. Moreover, temporal fact extraction has been verified as a more precise method for exploring time-related facts. One recent work presented TFWIN [5] platform on this topic. They first used MetaPAD textual pattern mining technique to generate patterns, entities, and values, then use a temporal tagger to find time indicators in the text. An example is given to find one country's presidents and their office term represented as $(e, v, [ts, te])$, like Mexico, Vicente Fox, [2000, 2006]. This fact can be extracted from multiple patterns, such as $Country president $Person. Based on this, they applied constraints learning approach to estimate pattern reliability and tuple reliability. And make them enhance each other in every iteration to achieve the final result.

However, the temporal information they extracted only accurate to year. In the real world, one country could have two presidents in one year. For example, the office term of Vicente Fox is from December 2000 to November 2006. After him, Felipe Calderon served as president of Mexico from December 2006 to November 2012. If we talk about the president of Mexico in 2006, should we say Vicente Fox or Felipe Calderon? This reveals that accurate temporal fact to year is not enough. Therefore, we proposed Misread-Proof Temporal Fact Extraction (MPTF) method which accurate time related extracted information to *month*, to offer a more explicit and reliable results. Our main contributions in this paper are listed as follows:

- We develop the MPTF method that extracts misread-proof temporal facts.
- Experiments on a large real-world dataset demonstrate the effectiveness and efficiency of our proposed method.

2 RELATED WORK

...

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 PROBLEM DEFINITION

Given a massive text source, can we extract trustworthiness temporal facts accurate to month.

4 APPROACH

Apply TFWIN[5] as platform.

5 EXPERIMENTS

In this section experiments proceses are introduced, results are discussed.

5.1 Experimental Setting

Data Set The dataset we collected rescources from English Gigaword Fourth Edition LDC2009T13 [51]. News articles here are published over the period mid-1990s to 2010. Six distinct international English newswire are involved, including Agence France Presse, Associated Press Worldstream, Central News Agency of Taiwanm Los Angeles Times/Washington Post, New York Times, and Xinhua News Angency. The size of entire dataset is 26,348MB (25.7GB), with 9.9 million articles and 4.0 billion words.

Evaluation We collected grouth truth, i.e., a set of true temporal fact tuples, on country's president from Wikipedia. The ground truth has 365 $(e, v, [ts, te])$-tuples of 130 countries, which accurate to month. We evaluate the performance of our method on mining the ground truth using standard Information Retrieval metrics: precision, recall, F1 measure and AUC (Area Under the Curve). For all of the metrics, the higher scores indicate that the method has better performance.

5.2 Experimental Results

6 CONCLUSIONS

Conclusions.

REFERENCES

[1] Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. MetaPAD: Meta pattern discovery from massive text corpora. In *KDD*.
[2] Yukun Ma, Erik Cambria, and Sa Gao. 2016. Label Embedding for Zero-shot Fine-grained Named Entity Typing.. In *COLING*. 171–180.
[3] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *EMNLP*.
[4] Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 995–1004.
[5] Xueying Wang, Qi Li, and Meng Jiang. 2018. On the Power of theWorldâĂŹs Invariants: When Truth Finding Meets Temporal Unstructured Data. In *KDD*. processing.