UTILISING PARAGRAPH VECTORS AND GAMIFIED MOBILE APPLICATION METRICS TO MEASURE TEAM EFFECTIVENESS

A Dissertation

Submitted to the Faculty

of

City, University of London

by

Edward Atkins

In Partial Fulfillment of the

Requirements for the Degree

of

MSc. Data Science

January 6 2017

City, University of London

London, United Kingdom

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed:

To my parents, Alison & Charlie, for supporting this new chapter of my life.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Atkins, Edward MSc, City, University of London, January 6 2017. Utilising paragraph vectors and gamified mobile application metrics to measure Team Effectiveness. Academic Supervisor: Tillman Weyde.

This project was undertaken as part of an internship with Know Your Crew, a gamified mobile platform that is designed to help teams connect and build trust, and gives managers actionable analytics to optimize team dynamics. The purpose of this project was two-fold, to develop a sophisticated language model for use in a variety of machine learning tasks, and to use this model, as well as other metrics from the companies gamified mobile application to predict team effectiveness. After training and evaluation, the Doc2Vec language model implemented produced strong results for document query tasks, and outperformed benchmark alternatives in sentiment analysis tasks. When this model was used to assess the relationship between application outputs and team effectiveness, it was found that there is a strong relationship between the similarity of answers provided by team-mates to personal-focused questions and their self-reported levels of team effectiveness.

1. INTRODUCTION AND OBJECTIVES

This report is based on the work undertaken as part of my internship for Know Your Crew. Know Your Crew is a gamified mobile platform that is designed to help teams connect and build trust, and gives managers actionable analytics to optimize team dynamics.

The core functionality of the Know Your Crew mobile application provides users within a workplace team a set of five questions weekly. There are five different question types within the application, but each requires a free-text response from the user. Within each team, players are divided into smaller groups (e.g. a team of 20 may be grouped into five teams of four). When each player within one of these smaller groups has completed the question set, each player is asked to predict which other player offered what answer, with a leader board shown on completion to indicate the most accurate player in prediction. Figure 1.1 and Appendix B provide a visual demonstration of this game-play.

My work with the company has been to develop an analytics pipeline to feed into Know Your Crew's web portal which provide managers a team snapshot across several key metrics that summarise responses and other activity within the mobile application.

This first iteration of the portal is largely descriptive, processing and condensing the data from the mobile application into a digestible form that gives managers insights into the view of their team on different topics related to both work and life. Ultimately however, Know Your Crew desires to provide predictive analytics services to its enterprise customers. One such proposed areas was to use data collected within the application to assess the effectiveness of a team.

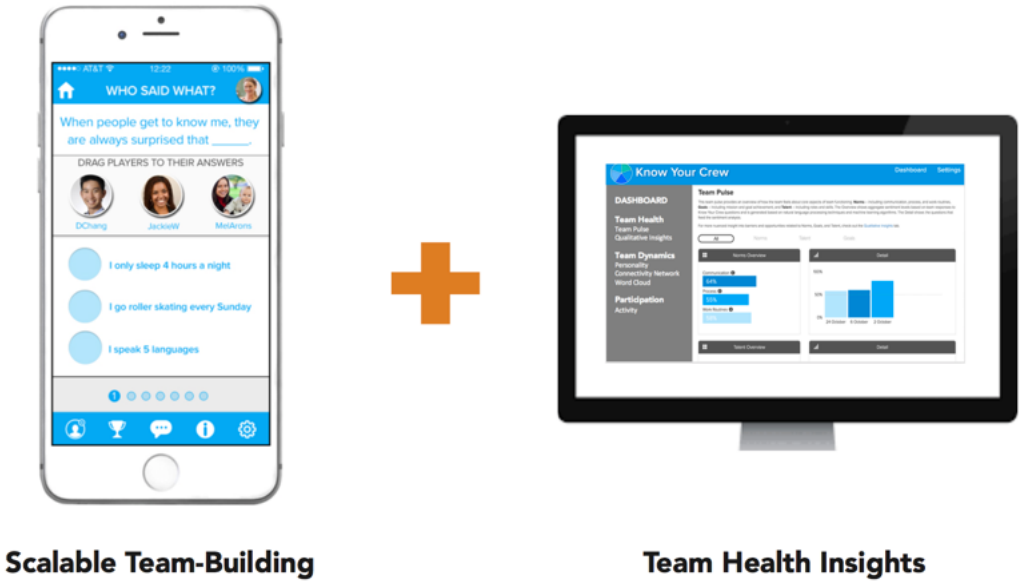


Fig. 1.1. Know Your Crew application

The primary source of data collected by the application are the responses provided by the users to each question. The free text format allows responses to be open ended and not necessarily constrained to a limited range of answers (e.g. multi-choice).

There are obvious challenges to this data format; in that it requires a level of sophisticated Natural Language Processing (NLP) to interpret and generate meaningful insight from the textual answers. But at the same time this format also creates significant opportunities because it allows users to respond in any manner they want, providing nuance to their answers, and exposing the analysis to their unique linguistic style. With the correct NLP techniques and machine learning algorithms, the Know Your Crew dataset has the potential to be a very powerful source for team-focused predictive analysis.

The base language model developed for this project was an implementation of the popular Doc2Vec algorithm, which can 'infer' a fixed-length vector from a variable

length document. The vectors generated by this algorithm have been proven to be particularly useful in a range of machine learning and other NLP tasks.

Aside from features generated employing the Doc2Vec language model, other features were generated from players' interaction with the Know Your Crew application. Specifically, these were prediction accuracy, engagement with the application and pro-noun usage.

A team effectiveness survey issued to class-based teams at the Wharton Business School, a top-ranked East-Coast USA university, was used to correlate these metrics with the self-reported effectiveness of each team.

There were two primary goals of this project undertaken for your Know Your Crew, which define the following analysis, and the overall structure of the report.

- Implement a language model (i.e. the Doc2Vec algorithm) that can be used by the business for a variety of natural language machine learning tasks
- Assess whether different metrics that can be extracted from the mobile application, are predictive of the different elements of 'Team Effectiveness'

2. CONTEXT

2.1 Overview

The critical context and literature review for this report has been divided into two sections in line with objectives of the project outlined in Section 1 above. This section of the report looks to provide the theoretical and literature based context that informed the development of the two pieces of analysis.

The first part explores the literature of language models that influenced the model selection and implementation choices for this project. The second part investigates the research regarding team effectiveness, including the definitional issues raised, the components that comprise it, and the work place factors that can affect it and how it might be estimated and/or predicted.

2.2 Language Models

2.2.1 The Distributional Hypothesis

The simplest Natural Language Processing (NLP) techniques treat words as "discrete atomic symbols" [1]. For example, machine learning systems that utilize the popular 'bag-of-words' method encode each word's term frequency in an unordered vector. Employing this method makes it difficult to capture the meaningful and often complex relationships that exists between words, for example negations constructions like "not sad" [2], the Capital City - Country relationship between the words 'Paris' and 'France', or the common characteristics shared by the words 'Banana' and 'Orange' (i.e. primarily that they are both types of fruit). Similar models, which look to assign probabilities to a string of words, such as n-Gram models [3] also face these limitations. These models are all forms of 'local representation' [4], where each word

or n-gram is represented by a single neuron in the model in a one-to-one relationship. This locality ensures that these models will fail to capture the componential elements of language, where words share some, but not all characteristics. For example, an orange and a banana both share the characteristic of being a fruit, but they do not share other characteristics like colour or shape.

More sophisticated NLP models have utilised Harris' (1954) "Distributional Hypothesis" [5], that proposes that it is the context in which a word resides that offers clues to its meaning. As described by Firth (1954) [6], words are "characterized by the company they keep".

In models that utilise local representation, each point in the model's sparse feature vector represents a single word (or n-gram). Conversely in a distributed representation, 'many-to-many' relationships are defined - each point (neuron) in the vector contributes to the representation of many (or all) words, and each word is represented by many words. One of the many benefits of representing a vocabulary through such a multi-dimensional distribution is to combat the 'curse of dimensionality'that hampers locally representative models [7]. The primary feature of these distributed models is that each word (or n-gram) in a vocabulary can be represented by a multi-dimensional vector, these representations are also known as 'word embeddings'.

Because they can represent ('embed') a vocabulary in a continuous vector space, these models are known as Vector Space Models (VSM). A feature of these models is that semantically similar words are mapped to nearby points [1]. In a well-trained model, it would be expected that the words 'orange' and 'banana' would sit relatively close in the vector space due to these two objects' shared componential properties.

2.2.2 Word2Vec

One of the most popular implementations to learn vector representations of a vocabulary is the Word2Vec model [8]. Although it is commonly referenced as one entity, Word2Vec describes two distinct training algorithms, the Continuous Bag-

Of-Words (CBOW) model, and the Skip-Gram model. The fundamental difference between the two algorithms is that CBOW predicts target words from its context (i.e. its neighbouring words), while inversely Skip-Gram predicts the context words from the singular target word. Both models train a one hidden layer neural network, typically with stochastic gradient descent and backpropagation [9], to predict words using a softmax classifier. In practice, the efficiency of this model is improved by classifying a more limited set of words by using hierarchical softmax or negative sampling techniques.

Post-training, each word in the vocabulary can be associated with a vector of synapses connecting the input layer to the hidden layer. This vector is the 'Vec' of the Word2Vec model; it represents the word in the multi-dimensional space of the entire vocabulary. Not only does this model produce a vocabulary space in which similar words are close together, the relative position of words captures useful semantic relationships.

For example, it has been shown in a vector space generated by a well-trained model with sufficient vocabulary; that the algebraic operation vector('King') - vector('Man') + vector('Woman') results in a vector that is most similar to the vector for the word 'Queen' [8].

While such relationships captured by the Word2Vec model are interesting, attempts to utilize their outputs for machine learning tasks such as sentiment analysis and document query have proved less fruitful. Machine learning algorithms typically require fixed length-inputs, meaning that n-word documents require an intermediary step to convert each of the n-vector outputs to a singular representation. One of the more popular methods has been to take the weighted average of all words in the document, though intuitively as the document length grows (e.g. a movie review) this approach is problematic. For classification, this method has shown to be no more effective than uni-gram and bi-gram Bag-Of-Words models [10].

2.2.3 Doc2Vec

Le and Mikolov (2014) [10] present a model that alleviates the problems associated with utilizing the Word2Vec model for machine learning purposes. The 'Paragraph Vector' model (a.k.a. Doc2Vec) is designed to produce a single vector representation for a paragraph or document of variable length.

The training algorithm is similar to that of the Word2Vec model (including the utilisation of stochastic gradient descent and back propagation but with the addition of a Document ID whilst predicting a target word or its context.) The Document ID plays the same functional role as an additional word in the target word's context. The vector for each Document ID is shared across all contexts in a document but is not shared across documents.

When using the Word2Vec model for machine learning tasks, it is limited to the words in its training vocabulary. With a training corpus of a sufficient size, this is unlikely to be problematic in application. However utilising the Doc2Vec model, the exponential combination of words in any document means that it is highly improbably that a new document is identical to one seen in training. Hence, the Doc2Vec model requires an additional step to calculate the vector for an unseen document, an 'inference step'. In this step the parameters for the rest of the model, the word vectors and softmax weights are fixed, and the document vector is obtained using gradient descent. The inferred vector can then be used as a representation of that document, for use in machine learning tasks such as sentiment analysis and search queries.

2.2.4 Gensim implementation

To generate a Doc2Vec model for this project, the popular Gensim library was used [11]. This library is particularly well known for its efficiency, because it utilises the Cython library, an 'optimising static compiler' that gives the Python code C-like performance. The following section discusses the Doc2Vec algorithm in relation to

this implementation, and the relevant literature. This is done by in turn focusing on each of the hyper-parameter choices provided by the Gensim library.

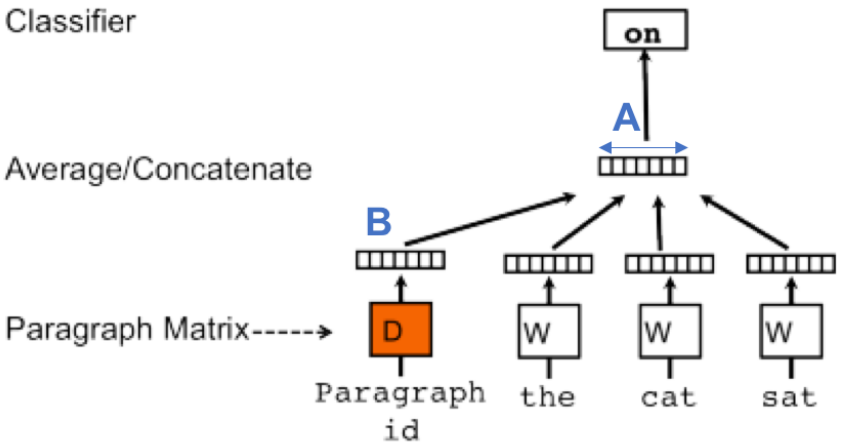


Fig. 2.1. Doc2Vec Distributed Memory Model

Distributed Memory or Distributed Bag of Words

This parameter choice defines the Doc2Vec model's training algorithm - like the Word2Vec model there are two distinct choices that can be utilised, either the 'Distributed Memory (DM)' or the 'Distributed Bag of Words (DBOW)' algorithm. These two algorithms are roughly analogous to the 'Continuous Bag of Words' and 'Skip-Gram' alternatives in the Word2Vec algorithm respectively. Figure 2.1 shows the DM model, while Figure 2.2 shows the DBOW model.

The DM is comparable with the Word2Vec model, with the paragraph vector effectively treated as another word (see Label B in Figure 2.1). The word vectors within a window and the paragraph vector are combined to predict the target word. By comparison, the DBOW model ignores the context of words and instead tries to pick words randomly sampled from the paragraph.

Distributed Memory - Concatenation

This parameter choice applies only to the DM model. The default setting in the Gensim implementation is to sum all context word and paragraph vectors to predict the target word. This alternative instead concatenates all the vectors, resulting in a much larger vector, and therefore a significantly greater training time.

Distributed Memory - Mean

As with the Concatenation setting described above, this parameter only applies to the DM model. Instead of summing (or concatenating) word and paragraph vectors the mean is taken.

Size

The size parameter defines the number of neurons in the hidden layer of the Doc2Vec neural network. This defines the dimensionality of the feature vector that represents not only each document, but each word in the vocabulary. The default dimensionality in the Gensim library is 300, while the original paper used 400 dimensions for their experimental protocol [10]. Other illustrated examples found online typically used in the 100 to 600 range for the vector dimensionality. In Figure 2.1 the size of the vector is shown by Label A.

Window

The context window defines the maximum distance between the target word and any context word. Words that fall within the context window are summed, average or concatenated with the paragraph vector to predict the target word (in the case of the DM algorithm). Because the vocabulary is pruned, the context window is effectively not fixed in size.

Previous work in relation to the context window sizes effect on the Word2Vec algorithm found that larger windows are effective in capturing topic information - determining what words are used in similar documents. While smaller windows were more effective in assessing functionally similar words, for example synonymous parts of the vocabulary. [12]. However, this work did not provide any guidance as to the implications this might have for the Doc2Vec algorithm.

Minimum Count

The minimum count defines the minimum frequency for a word to be included in the model's vocabulary. The choice of this variable is a trade-off between effectiveness and efficiency. In terms of performance on an unseen validation task, a model that includes a greater vocabulary should perform better ceteris paribus. However, the more marginal (less frequent in the given document corpora) the word, the more marginal the gains of performance, and the greater cost of using the model in production in terms of processing time. Note: the Gensim implementation also includes a parameter to use the maximum vocabulary size which effectively performs the same function.

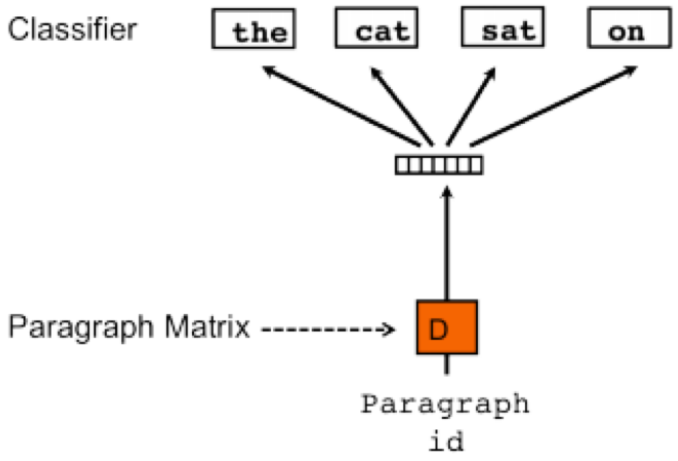


Fig. 2.2. Doc2Vec Distributed Bag of Words Model

Alpha

Alpha is the learning rate of the network and will drop to zero over the training iterations.

Hierarchical Softmax

Because of the size of the vocabulary created by the Doc2Vec model, instead of calculating probabilities across the whole softmax layer (which is the size of the

models vocabulary), it can be more efficient to construct a binary tree [13]. Instead of predicting and normalizing across all words in the vocabulary, this process is reduced to a small number of nodes in the tree. The Gensim implementation uses a binary tree called a Huffman tree where the most frequent words have the shortest paths.

Negative Sampling

An alternative to Hierarchical Softmax is Negative Sampling. The method is a simplification of Noise Contrastive Estimation (NCE) applied to language modeling by Mnih and Teh [14] that uses logistic regression to distinguish target words from 'noise' words using logistic regression.

The task becomes to distinguish the target word from k randomly drawn negative samples from the vocabulary [8]. The Gensim implementation of the Doc2Vec algorithm allows scope to combine both Hierarchical Softmax and Negative Sampling in the training algorithm.

2.3 Team effectiveness

The second objective of this report was to assess the extent to which metrics generated by the Know Your Crew application could be used to predict the effectiveness of a work place team. For this purpose, there were two primary perspectives from which to examine the literature.

The first was to understand the definitional issues that permeate the team effectiveness literature. How is the term 'team effectiveness' commonly defined, what are the different components of the workplace that are thought to comprise it, and how are these measured?

The second perspective is to understand research that has sought to provide correlative links between measures of team effectiveness and other measurable quantities, particularly ones that are relevant to the kind of outputs provided by the Know Your Crew mobile application.

2.3.1 Defining team effectiveness and its components

Teams, and their effectiveness, are a crucial element of the modern workplace. Team effectiveness theory has traditionally been posited in terms of an input-process-output (IPO) framework [15]. Input refer to the characteristics and resources available to a team at an individual, team and organizational level. Processes are the activities in which teams employ their combined resources to resolve the tasks assigned to the team.

Output is the crucial component of this framework in thinking about the measurement of team effectiveness. Phillips (2012) [16] defines it purely in terms of the output or work quality of team. While Hackman (1987) [17] defines the output component in terms of three facets - the performance of the team judged by external sources, the extent to which individuals need are met, and the viability or the willingness of members to remain within the team. Druskat (2001) [18] also views team effectiveness through three components - the trust amongst group members, a sense of efficacy and a sense of group identity. While Luthans (2011) [19] cites five different components of team effectiveness that should be monitored.

2.3.2 Measures of team effectiveness used by Know Your Crew

Across the literature, there is a lack of a clear and consistent definitional approach to the issue of defining team effectiveness. This allowed some flexibility in designing a custom set of team effectiveness characteristics for this report. These characteristics are founded in the academic literature, but were selected in consultation with the Know Your Crew team to be most relevant to the core product and the different elements of team effectiveness that the application is intended to impact. Additionally, only those measures that could be reliably captured through self-reporting were chosen, as this was the mechanism to be used in data collection for this report.

These different elements of team effectiveness are discussed below - with reference to both the supporting evidence in the academic literature, as well as their relevance to the Know Your Crew application.

Cohesion

Probably the most common feature across the literature defining team effectiveness, is the inclusion of a component analogous to 'team cohesion'. Cohesion being the shared bond that drives team members to want to stay together and remain united in the pursuit of a common goal [20]. A measure similarly defined to team cohesion, 'bridging social capital' was assessed in a study of Facebook users [21]. Bridging social capital is defined as the "building of connections between heterogeneous groups" [22]. The Facebook study is demonstrative in the context of this analysis, because it illustrates the ability to link in-application metric to a core component of team effectiveness.

Trust

Trust is the willingness of a person to make themselves vulnerable to the actions of another based on the belief or expectation that the other person will act in a manner favourable to the individual [23]. Druskett (2001) [18] outlines trust as one of the three key elements of an effective team, alongside a sense of efficacy and a sense of group identity, that are empirically linked to high performance [24]. Trust is important to an effective team because it encourages risk taking [25], facilitates information sharing [26], encourages collaboration [27] and enhances productivity [28]. The Know Your Crew application specifically targets trust by providing an open and non-threatening environment for workers to share their thoughts and opinions amongst their teammates.

Perspective Taking

Perspective taking is the ability to understand a situation from another persons point of view, including their thoughts, feelings, desires, motivations and intentions [29]. Perspective taking has been shown to reduce prejudice and stereotyping [30], and is therefore a crucial element of team effectiveness for its ability to unlock the

benefits of workplace diversity [31]. Perspective taking is viewed as one of the core elements of the Know Your Crew application, as the prediction component of the game play actively engages to the user to consider their teammates' thoughts across a wide range of work related and non-work related topics.

Role Clarity

Role clarity is defined as the state where an individual in a team has clear and sufficient about their role within that team [32]. The evidence as to its benefits for team efficacy are conflicting. Erikson (2012) [33] found that role clarity is crucial to effect high performing teams, allowing them to overcome barriers such as location, lack of a common language and tasks with highly complex and diverse components. However in contrast Lynn (2015) has found that while the clarity of a teams vision is import to the teams efficacy, there is no statistically significant link between role clarity and team performance [34]. The Know Your Crew application seeks to improve role clarity specifically by exposing game players to questions that highlight each persons role in the team, and whether players feel comfortable with their individual level of role clarity.

Communication

Pentland (2012) [35] found patterns of communication to be the most important predictor of a teams success. Communication was framed in terms of three components - the quality and quantity of the communication (the 'energy'), which specific team members were communication well with each other (the 'engagement'), and the teams communication with outsiders (the 'exploration'). Know Your Crew hopes to improve communication in the work place by providing teammates with insight into each other's thoughts and opinions, and providing a common set of topics on which to engage.

Creativity

This measure was defined as where teams are willing to try new different approaches to their and otherwise look to improve existing tasks. Creativity is considered the first step of innovation [36] and has been described as the cornerstone

of organizational change, the foundation of innovation, and a key to organizational effectiveness [37].

2.3.3 Measures predictive of team effectiveness

The metrics extracted from the Know Your Crew application are in most part unique to that context - so it is not possible to borrow directly from the literature to implement different measures predictive of team effectiveness. However, the different ways in which this data has been used to generate independent features in this report has drawn inspiration from different areas of the research that have proven correlation between measurable elements of work place groups and their effectiveness as teams (or analogous measures). Below is discussed some of the most relevant research, and its potential relevance for generating independent variables from the Know Your Crew data.

Diversity

Diversity within the workplace has been proven to be associated with the effectiveness of teams when compared to work groups that are homogeneous. Diversity has a particularly large impact in the areas of problem-solving, conflict resolution and creativity [38]. These effects are prevalent across areas such as race, ethnicity, gender and sexual orientation groups [39]. It was hypothesised that this diversity could be borne out in diversity of response to Know Your Crew questions, and therefore such a measure may be a useful predictor of team effectiveness.

Pronouns

A common test, popularised by former Secretary of Labor Robert Reich, proposes that when workers are asked questions about their own company, their use of pronouns is indicative of the company's success [40]. Companies whose workers commonly describe the organisation in terms of 'we' or 'us' as opposed to 'they' are believed to have the markers of success. These features are something that can be directly

measured in a user's responses to the Know Your Crew application, and was therefore considered a suitable measure to assess against team effectiveness.

Language Style Matching

There is a wealth of research that investigates the relationship between social mimicry and the prevalence of positive social dynamics between groups that undertake such behaviour. Most importantly, mimicry has been linked to team performance [15]. Particularly relevant to the this report, is that similarity of non-verbal language used by people in communication, has proven to be predictive of their likely personal connection [41] [42]. It was hypothesised that evidence of language similarity driven by close team connections might be measured in the responses to certain Know Your Crew questions and therefore used as a predictor of team effectiveness.

3. METHODS

3.1 Overview

Like previous sections, this report's methodology is structured in terms of its two primary objectives. First it describes the training and evaluation of a Doc2Vec model, to be utilised by Know Your Crew for a variety of NLP and machine learning tasks. Secondly it describes the process to assess whether there is correlation between metrics generated by a user's interaction with the Know Your Crew application and self-reported measures of team effectiveness. The report methodology, including the intersection of the two report objectives is summarised in Figure 3.1.

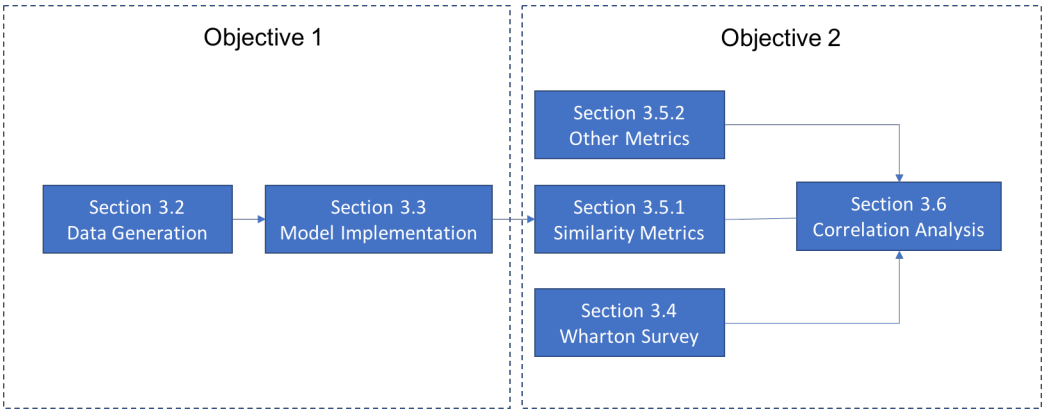


Fig. 3.1. Method Summary

3.2 Data generation

Unlike many other distributed representation models in NLP, the Doc2Vec (and Word2Vec) algorithms are neural networks with just a single hidden layer. For this

reason, the data requirements to train an effective model are significant. These are not deep learning models, instead they are designed to trade complexity for efficiency in training on large text corpora [8]. Commonly, the algorithm is trained on a data-set that is tens of gigabytes, comprising publicly available sources such as the latest Wikipedia dump [43] or the 'One Billion Word Language Modelling Benchmark' [44].

The Know Your Crew data-set at the time of this report contained over 13,000 responses. Whilst large in some respects, it was considered unlikely that this data-set would be sufficient to train an effective Doc2Vec model. Therefore the proprietary data-set was supplemented by a Twitter data set of over 600,000 tweets and the latest Wikipedia dump which had over 4,000,000 articles.

The Gensim Doc2Vec model requires data in a plain text format, with each line representing a new 'document'. Below is a brief summary of the processes undertaken to parse and transform each source to this format.

Wikipedia

The Gensim library includes a module that coverts the xml Wikipedia dump to the required plain text format. Additionally, llematization was considered as an extra pre-processing step to reduce the overall vocabulary size and improve the efficiency of the trained model. However a test of Gensim's in-built llematization method on the Wikipedia corpus revealed that the significant increase in processing time was unlikely to be worth the trade-off.

Twitter

The 'tweepy' Python library was used to access the Twitter API. To focus the results of the search, the API was used to search specifically for tweets containing words that were most used by Know Your Crew enterprise users (excluding stop words). The tweets returned by this process were then cleaned to remove hyper-links, 'mentions' and 'retweets' and tweets of length less than five words. Tweets were also filtered employing a Python library 'guess-language' to exclude those results that were not English. The scraped tweets were saved to a local MySQL database using the 'SQLAlchemy' library, and subsequently read into a plain text file for model training.

Know Your Crew

For model training and testing purposes, the Know Your Crew database was provided as a daily MySQL dump. The 'pymysql' library was used to extract answers from the database, before filtering based on question type (i.e. certain question types were excluded) and ensuring that the sentence length was at least five words, before being added to a plain text file.

3.3 Model implementation

The Doc2Vec model is an unsupervised algorithm, and as such there is no explicit evaluation method. The original Doc2Vec paper [8] presents two form of evaluation. The first is to use the model to conduct a sentiment analysis on a benchmark data-set (e.g. the IMDB data-set). The second is to employ the model in a search query task, using the vector representation for a 'search document / query' to find the most similar document from a selection of documents. In this task one of these documents is drawn from a set of similar document, whilst the others are drawn randomly from some mutually exclusive set.

3.3.1 Model selection using a query task

To optimise for hyper-parameters in the Doc2Vec model the search query task was preferred for this report - this choice was made due to a couple of factors. Firstly, it is reliant solely on the outputs of the model, and does not require consideration of a secondary machine learning algorithm (e.g. a sentiment analysis model may utilise a Logistic Regression or Support Vector Machine to classify documents). Secondly it was decided that a document similarity task would be more relevant in the context of future Know Your Crew uses of the model.

Two custom validation data-sets were created based on real answers from the Know Your Crew database. The first included 191 responses given to 14 different questions, while the second included 135 responses based on 9 different questions.

Similar answers were identified and grouped for a given question. Subsequently, for each similar answer pair, a 'test query' was created by sampling two random (but non-similar) responses to the same question. In total 535 of these test queries were created for the first validation set, and 1018 for the second set.

For each test one of the similar responses was chosen at random as the 'query term' and one as the 'response term', and then the Doc2Vec model used to infer a representative vector for all four documents. Employing the dot product of two vectors as the measure of similarity, the test query was judged successful if the dot product between the query response and the test response was greater than between the query response and either of the randomly sampled responses.

Figure 3.2 illustrates an example test query, in response to the question, "If we could add one skillset to our skillset to our team, what would it be?". In this case, the query responses and test responses were judged in the same category because they both refer to 'analytical' capacity, while the randomly drawn responses refer to 'UX' and 'meeting time management' respectively. Given the similarity measures depicted in the figure, this particular example would have been assessed as correct ($0.5 > 0.2 > 0.1$).

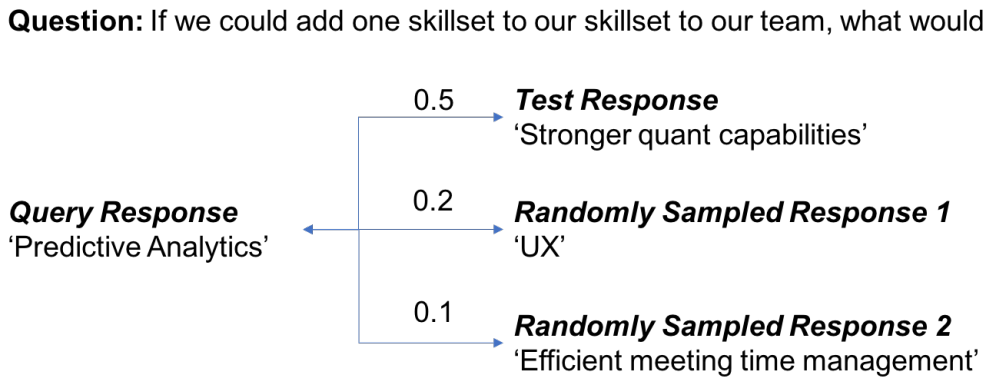


Fig. 3.2. Doc2Vec Test Query

In evaluating the model, there were two dimensions of model selection explored. The first was the data-set used, the optimal combination of the Wikipedia, Know Your Crew and Twitter data-sets described previously in Section 3.2. The second was the different hyper-parameter combinations allowed by the Gensim implementation of the Doc2Vec algorithm.

A complete model optimisation process would involve a full grid search across all possible hyper-parameter, and data-set, combinations. However due to the size of the data-set, and the subsequent training time, this would not prove feasible within a time frame suitable for this report. For the full data-set a training iteration would take between 6 and 20 hours, depending on the specific hyper-parameter combination. Assuming (conservatively) only three choices for each parameter, there would be over 60,000 different models to evaluate in a full grid search.

Three steps were taken to adapt a full hyper-parameter grid search to a process appropriate for this report. Firstly, the data-set combination problem was separated from the hyper-parameter search. Secondly, for the initial hyper-parameter search, the data-set was reduced to 1/10 of its original size, by taking every 10th document. Finally, the hyper-parameter search borrows from the greedy hill-climbing algorithm method described by Minah (2015) [45]. The algorithm searches across a whole single dimension of parameters and picks the one which increases the value of the objective function (in this case, the accuracy on the evaluation set) by the most. Each parameter is visited in turn, until no parameter change increases the value of the objective function. This basic approach is problematic in that the final results are biased by the point of initialisation of each hyper-parameter, as well as the order in which the hyper-parameters are searched - the final parameter set determined may be a local rather than a global maxima. To mitigate this issue, the search process was repeated three times on the first validation set, with both initialisation and order randomised. The three resulting hyper-parameter sets, were then used to train Doc2Vec models on the whole data-set, and tested against the second validation set.

The best performing of these three hyper-parameter combinations was used for the succeeding analysis.

To further optimize this process, an element of subjectivity was layered to ignore hyper-parameter choices on subsequent passes over the set of hyper-parameters, that while valid were obviously deleterious to the performance of the model.

3.3.2 Validation on sentiment analysis task

As a separate validation step, the Doc2Vec model was also evaluated on a sentiment analysis task - like the original Paragraph Vector paper [10]. That paper used a benchmark data-set, the Stanford Sentiment Treebank Dataset [46], to compare the performance of their model against several alternatives, including a bag of words model, bi-gram model, word vector averaging model and a Recursive Neural Tensor Network.

The models were evaluated against both a coarse-grained and fine-grained sentiment labelling system. The coarse-grained system included just positive / negative labels for the documents in the data-set, while the fine-grained system outlined five different categories; Very Negative, Negative, Neutral, Positive, Very Positive.

For this report, the Doc2Vec model was compared in a sentiment analysis task against both a bag of words and n-gram model. In terms of a classification system, a middle ground was chosen between the coarse and fine-grained systems described above, with answers labelled Negative, Neutral or Positive.

Instead of evaluating on a benchmark data-set, this analysis focused on a sentiment data-set developed from actual Know Your Crew responses. Over 400 responses to questions tagged by the Know Your Crew team as sentiment relevant were hand-labelled into one of the three categories. This decision to use this over a benchmark data-set for validation was driven by the intention to deploy a sentiment analysis model within the Know Your Crew model using this same data.

A 10-fold cross validation process was used to evaluate a logistic regression model trained on the vectors inferred by the Doc2Vec model from classified responses. The bag-of-words and n-gram models were also trained using a logistic regression algorithm.

3.4 Wharton Business School survey

To collect data relating to team effectiveness, Know Your Crew partnered with the Wharton Business School. The Know Your Crew mobile application was provided to the Management 240/740: Leading Effective Teams class over the course of the Fall 2016 semester, with participating students asked to complete two surveys during the semester to assess their self-reported measures of team effectiveness. The class size was 40 students, comprising both undergraduate and graduate students.

Coursework for this class was undertaken in teams of four, and the assumption was made that these would be adequate proxies for workplace teams. At the beginning of the semester, each team was asked to download the Know Your Crew application, and over the course of the semester they were issued five questions weekly; with the game to played within their coursework groups. The class were also issued two identical surveys over the course of the semester using Google Forms, the first in Week 3 of semester and the second in Week 10, with the results of the latter to be used for the team effectiveness correlation analysis. Design and analysis of the survey were managed by this report's author, while communication and issuance was undertaken by other members of the Know Your Crew team.

It was initially intended that this research would also assess any impact of using the Know Your Application on team effectiveness over time. However this was ultimately not possible, due to certain factors that made it impossible to randomly separate a control group from the Wharton cohort.

3.4.1 Survey design

Although all responses were to be kept anonymous, identifying details including a person's full name and team name were collected to allow pairing of survey responses and Know Your Crew application data. Besides this information, six questions were asked of respondents in the survey, each corresponding to a component of team effectiveness defined in Section 2.3.2. Taking a multi-dimensional approach to assessing team performance is supported by a meta-analysis of studies into team cohesion conducted by Salas (2015) [47]. This research found that researchers should adopt multidimensional definitions, as "multidimensional conceptualizations found significant cohesion-performance relationships more frequently (69%) compared to unidimensional conceptualizations [57%]) ".

The questions developed for this survey, were designed in the language employed by a similar study of team effectiveness of student team effectiveness conducted at MIT. [48] The survey asked the following questions - with responses allowed on a gradient scale from 1 to 7.

- The degree of trust amongst team members [Trust]
- The clarity of roles on your team [Role Clarity]
- The degree to which your team enjoys working together [Cohesion]
- The degree to which communication between team members is open and participatory [Communication]
- The degree to which your teammates can see things from your point of view [Perspective]
- The team's ability to try new or different ways of doing work together [Creativity].

3.5 In-application metrics

In total eight different independent variables derived from the Know Your Crew application were assessed for correlation with team effectiveness measures collected from the survey. To ensure that results were not skewed by questions answered by only a few participants, data was collected only for questions with a minimum 10 responses from the Wharton cohort.

3.5.1 Similarity-based independent variables

Four of these metrics were based on a similarity analysis of responses to specific question sets within the Know You Crew application. The similarity analysis assessed how semantically similar a players responses were to those of their teammates across these questions.

For each question (i) in a set of questions (n) answered by a given player (x), the Doc2Vec model was used to infer a representative vector (d) from their response (r). The model was also used to infer vectors for every other answer (y) in the set of teammate answers (m).

The dot product of the vectors for every response pair $d(r_{ix})$ and $d(r_{iy})$ is averaged to determine a similarity score for that question. This process is repeated and macro-averaged across all questions (n) in the category as defined in Equation 3.1 to produce the similarity score (sim).

$$sim_x = \frac{\sum_{i=1}^{n} \frac{\sum_{y=1}^{m} d(r_{ix}) \cdot d(r_{iy})}{m}}{n} \tag{3.1}$$

Questions sets were divided by the expected nature of the response into the four sections outlined below. This was achieved through a qualitative analysis of previous answers to these questions in the application. It was hypothesised that similarity on different question types, may be differently correlated to a measure of team effectiveness. Team diversity [39] and Language Style Matching [42] for example, are both

shown to be positively related to positive team relationships, but they potentially could have opposing effects on the similarity score depending on the question type.

Positively-slanted work questions (Work positive)

These were questions relating to the teams work place, where the player is asked to identify positive things about their workplace, or where it would be typically expected that similar answers generated by teammates would be a positive outcome

Example questions of this type:

- "How would you summarize our teams mission in 5 words or less?"
- "If our team achieves [blank] and nothing else, we'll be successful"

Negatively-slanted work questions (Work negative)

These were question relating to the teams work place, where the player is asked to identify problematic features of their workplace or areas for improvement. For these questions, an alignment of opinion might be actually be a negative outcome, because it signifies that there is an obvious issue that is identifiable by many of the team.

Example questions of this type:

- "If our team had a process for [blank] it would make our lives easier"
- "What would our team be like in perfect world"

'Most Likely To' questions (MLT)

These are question in the application in the form of, "Who is most likely to [blank]?". The players then have the option to auto-fill one of their teammates username as the answer. These questions are structured to be an opportunity for recognition within a workplace team.

Example questions of this type:

- "Who is most likely to have won a public speaking award we dont know about?"
- "Who is most likely to spot the pothole before we hit it?"

Personal questions (Personal)

Personal questions were identified as those questions that did not fall into the three categories above. Most often these questions would therefore ask the player to express an opinion or preference.

Example questions of this type:

- "If you ran a food truck, what would it serve?"
- "[blank]is my 'breakfast of champions"'

3.5.2 Other independent variables

Sentiment analysis questions (Sentiment)

Sentiment analysis, like the similarity metrics defined in Section 3.5.1 was conducted on a subset of questions within the application. In this case a set of questions were specifically designed by Know Your Crew to generate answers suitable for sentiment analysis. The Doc2Vec model trained and evaluated as described in Section 3.3 was used to generate the feature set for each response. The sentiment model used was developed in the same workflow outlined in Section 3.3.2. Although these questions and the sentiment model were created to assess sentiment for specific work place topics for this report these responses were aggregated to determine a measure of general sentiment towards the work place. The class probabilities output by the sentiment model were combined to create one sentiment score (sent), by summing the predicted probability of a positive statement (pos) and half the predicted probability of a neutral statement (neu). This scored was averaged across all sentiment answers provided by an individual as illustrated in Equation 3.2.

$$sent_x = \frac{\sum_{i=1}^{n} neu_{ix} * 0.5 + pos_{ix}}{n} \tag{3.2}$$

Example questions of this type:

- "[blank]: my feelings about our team's daily schedule"

- "How do you feel about our teams decision making process?"

Pronoun usage (Pronoun)

It has been argued that analysing the use of pronouns is an effective tool for assessing the health of an organization [40]. When describing their company or team, employees that prefer the use of third person pronouns are considered more likely to be disengaged than those who more frequently use first person pronouns. This metric was assessed only for work-related questions, where it was calculated the proportion of times that an individual used a first person plural pronoun ('we' or 'us') compared to a third person plural pronoun ('they' or 'them').

Application engagement (Engagement)

The engagement metric was determined by measuring the average number of the five weekly questions answered by a player through the semester.

Prediction accuracy (Accuracy)

Prediction accuracy measures the proportion of correct responses provided by each player in predicting their team-mates answers within the application. However, in some cases, not all players within a team would make it to the final prediction step of that weeks questions. This scenario had the potential to bias the prediction accuracy metric, because a smaller group would make prediction easier for the remaining players. To remedy this potential issue, prediction accuracy results were normalised for the number of players that made it through to the final prediction step.

3.6 Correlation analysis

Each of the in-application metrics detailed in Section 3.5 was calculated on a per-player basis. These were assessed against a combined team effectiveness score for that Wharton student, a summation of the six different measures surveyed. To determine the relationship between each of these independent variables and team effectiveness score, two different tests of correlation were utilised. Correlation is a bivariate analysis that measures the strengths of associations between two variables. Pearsons r

correlation is widely used to test the association between two linear variables. However, this test assumes that both variables should be normally distributed, and that any observed relationship between the variables will be linear in nature. An alternative is Spearmans rank correlation, a non-parametric test. The comparative strength of this test is that it is does make any assumptions about the distribution of the data, or the nature of the correlative relationship.

Both tests were used in this analysis to assess the correlation between team effectiveness and in-application metrics. Because there is a degree of inherit stochasticity within the Doc2Vec model, the analysis procedure was run several times and an average of the results taken.

4. RESULTS

4.1 Overview

This section provides an overview of the results of the analysis with regards to the methodology described in Section 3. It firstly describes the performance of the Doc2Vec algorithm, using a query test task to optimise across a set of hyper-parameter and data combinations, as well as the results of additionally validating the model on a sentiment analysis task. Secondly, it details the results of a correlation analysis, between metrics derived from the Know Your Crew application and the outcomes of the Wharton Business School survey.

4.2 Implementing the Doc2Vec algorithm

4.2.1 Data combination

The process of selecting an optimal combination of textual data sources was conducted preceding a search over the hyper-parameter space, as detailed in Section 3. From the three primary data sources (Know Your Crew, Twitter and Wikipedia), seven different combinations were generated and tested; each data source individually, the three different pair combinations, and the combination of all three data sources. In lieu of a pre-optimised hyper-parameter set, the default Doc2Vec settings of the Gensim implementation were utilised for this step.

The best performing combinations on the query test set were the combined Wikipedia/Know Your Crew data-set and the Wikipedia only data-set. In likelihood, the Know Your Crew data-set was of insufficient size to make an impact on the final results; however with the knowledge that this data-set would grow in the future (and the Doc2Vec

Know Your Crew	Wikipedia	Twitter	Accuracy
✓	✓	✓	0.65
✓	✓		0.72
✓		✓	0.54
	✓	✓	0.65
✓			0.35
	✓		0.72
		✓	0.53

Table 4.1
Data combination results

model re-trained utilising this report's workflow), the combined (Wikipedia / Know Your Crew) data-set was preferred for the subsequent analysis.

4.2.2 Hyper-parameter search

Subsequently, the Doc2Vec Model was trained and evaluated to optimise for its hyper-parameters, using the methods described in Section 3.3.1. Table 4.2 shows the resulting parameters resulting from the three hyper-parameter searches each with initialisation and order of search randomised, based on accuracy scores on the first of two evaluation data-sets.

Over the three experiments, the performance of the models ranged from 35-81% depending on the hyper-parameter combination selected. The best scores for each experiment, were 79%, 81% and 77% respectively.

One of the noteworthy outcomes of this testing process was that the use of only Negative Sampling (when not used in combination with Hierarchical Softmax) tended to produce results that were only barely better than that of random chance (i.e. 33%).

Parameter	Range	Best 1	Best 2	Best 3
Window	2-12	8	6	8
Iterations	4-12	10	10	10
Size	200-900	500	350	250
DM	0,1	0	0	0
DM Concat	0,1	0	0	0
DM Mean	0,1	0	0	0
Alpha	0.01-0.05	0.05	0.035	0.025
Hierarchical Softmax	0,1	1	1	1
Negative Sampling	0-20	0	0	0

Table 4.2
Hyper-parameter search results

4.2.3 Final model selection

These three possible hyper-parameter combinations were then tested on the second evaluation set, to determine which to use for the final Doc2Vec model. The scores of these models were, 69%, 70% and 67% respectively. As such, the second of these three models was used for the remaining analysis tasks.

4.2.4 Sentiment analysis

The final Doc2Vec model was additionally evaluated via a sentiment analysis task, using 410 responses taken from the Know Your Crew data-set. Each response was hand-labelled as either positive, neutral or negative.

The results from this task are shown in Table 4.3 - comparing the performance of a Bag of Words (uni-gram) model, an n-gram (n = 1-3) model and the Doc2Vec model across accuracy and F1 metrics, using a 10-fold cross validation process. All

Model	Accuracy	F1
Bag of Words	0.58	0.45
N-Gram	0.59	0.49
Doc2Vec	0.63	0.56

Table 4.3
Sentiment Analysis Validation.

models were trained on the three-class problem using a Logistic Regression model. On both measures, the Doc2Vec model significantly outperformed the other models based on these metrics. These results reflect the the original Doc2Vec paper [10] in demonstrating the superior performance of a Doc2Vec model in a sentiment analysis method when compared to other methods.

4.3 Team effectiveness

4.3.1 Survey summary statistics

In total, there were 40 participants from the Wharton cohort who undertook the 12-week Know Your Crew program. All 40 of these participants returned the preliminary survey at the beginning of semester and 31 returned the survey issued mid-way through the program. With the intention to replicate a work-place environment as much as possible, the latter survey was used for the correlation analysis, under the hypothesis that these groups would need some time to establish basic team dynamics. Table 4.4 illustrates the mean and standard deviations of each the seven team effectiveness measures for reach survey.

The initial purpose of issuing two surveys was to also assess whether use of the Know Your Crew application over time was effective in improving team effectiveness rather than just being predictive. This was ultimately not feasible, as the the class

	Survey 1	Survey 1	Survey 2	Survey 2
Metric	Mean	StdDev	Mean	StdDev
Trust	5.50	0.97	5.91	0.91
Role Clarity	4.50	1.20	5.38	1.08
Cohesion	5.53	0.93	6.0	0.97
Communication	5.94	0.83	6.09	0.84
Perspective	5.31	0.88	5.56	0.86
Creativity	5.34	1.16	5.5	1.0

Table 4.4
Wharton Survey Descriptive Statistics

Professor's reluctance to exclude any students from being able to use the Know Your Crew application, made it impossible to randomly separate a control group from the Wharton teams. However, it is still interesting to note the changes in the score throughout the semester.

All six elements of team effectiveness increased over the semester. The most significant of which was for the 'Role Clarity' question, which increased 19%. This finding is intuitive given the newness of the teams when the survey was first administered. Cohesion (9%), Trust (7%) and Perspective-taking (5%) also saw significant increases over these weeks. While Communication and Creativity (both 3%) both saw modest increases.

For the correlation analysis, and in line with the multi-dimensional approach recommended by Salas (2015) [47], these measures were combined to form an overall 'team effectiveness' score. The distribution of these scores is shown in Figure 4.1. Overall, the mean total team effectiveness score was 46.1 in the second survey, increasing from 43.4 in the first survey.

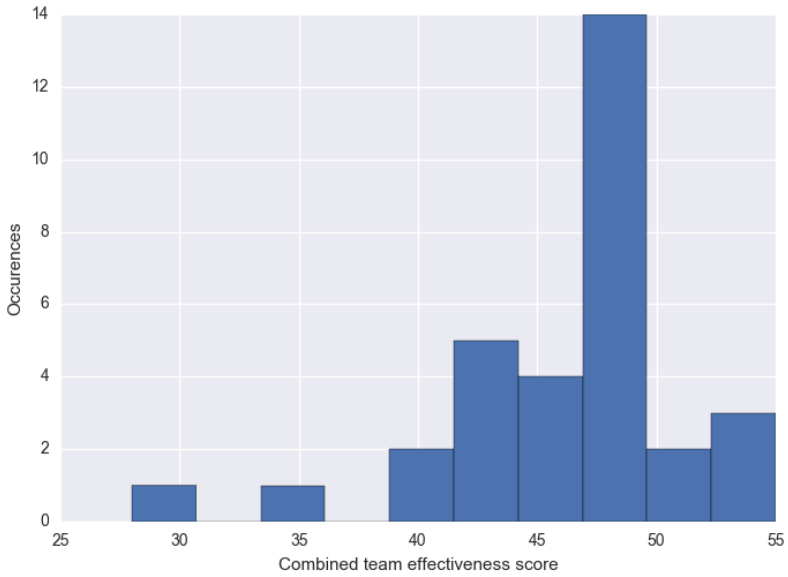


Fig. 4.1. Team Cohesion Score Histogram

The python library 'Scipy' was used to test the normality of the data. The test is based on D'Agostino and Pearson's [49] [50] test that "combines skew and kurtosis to produce an omnibus test of normality" [51]. The p-value for the combined survey results was $2 * 10^{-4}$, indicating that with a high degree of confidence we can reject the null hypothesis that the sample comes from a normal distribution. This illustrates the importance of considering both non-parametric as well as parametric measures in a correlation analysis.

4.3.2 Correlation analysis

The combined team effectiveness score was tested against each of the independent variables in turn to test for correlation, using both Spearmans and Pearsons tests of correlation. The results of this analysis are shown in Table 4.5 which indicates the correlation metric for both the Spearman and Pearson correlation scores as well as

the p-value for each correlation. The p-value in this case represents the probability of the result given the true correlation was zero (the null hypothesis).

Variable	Correlation Metric	Correlation	P-Value
Personal	Pearsons	0.467	0.016
Personal	Spearmans	0.513	0.007
MLT	Pearsons	-0.331	0.099
MLT	Spearmans	-0.223	0.273
Work Positive	Pearsons	-0.274	0.184
Work Positive	Spearmans	-0.072	0.733
Work Negative	Pearsons	-0.044	0.830
Work Negative	Spearmans	-0.276	0.172
Engagement	Pearsons	0.082	0.655
Engagement	Spearmans	0.011	0.952
Accuracy	Pearsons	-0.199	0.557
Accuracy	Spearmans	0.019	0.957
Sentiment	Pearsons	-0.016	0.938
Sentiment	Spearmans	-0.106	0.606

Table 4.5
Correlation Results

Of the eight independent variables tested, six showed no or weak correlation to team effectiveness. Another, personal pronoun usage, was excluded because of an insufficient use of these terms by Wharton players. The clear exception was the similarity of answer in response to personal questions, which showed a strong positive correlation with team effectiveness. It measured 0.467 using Pearsons correlation score and 0.513 on Spearmans, with p-values of 0.016 and 0.007 respectively. These results indicate that at a standard significance level of 0.05, we could reject the null hypothesis that there is no correlation.

Of the other six metrics tested for correlation, only the MLT measure showed potentially interesting results. Similarity of team response on MLT questions showed a negative correlation between similarity of answer and the team effectivness measure, -0.331 using Pearsons correlation and -0.223 using Spearmans correlation. This corresponded to p-values of 0.099 and 0.273 respectively. Whilst not statistically significant at the 5% level, these results are likely worthy of investigation in any future work. This result would mean that more diverse a response provided by a team on the MLT questions, the higher the level of self-reported team effectivenss. The other independent variables, prediction accuracy, application engagement, sentiment analysis and the similarity analysis for the positively-slanted and negatively-slanted work questions did not produce any correlation of note.

5. DISCUSSION

5.1 Overview

This section considers the results of the analysis considering the original objectives. It provides a discussion of these results, including their validity, potential applications and areas identified for improvement in any future work.

5.2 Doc2Vec Model

This project has produced a valuable tool and workflow for Know Your Crew to deal with sophisticated NLP tasks. After the hyper-parameter tuning process, the Doc2Vec model was performing above 81% and 70% on the query tasks outlined in the two validation sets. It also scored a significantly better prediction accuracy on the nascent sentiment analysis data-set than more the more traditional alternative of a bag of words model, producing an error rate of 32% when compared to the bag of words model of 52% and the n-gram model of 41%. The model is discussed below in relation to these illustrated results, including areas for improvement and its application within the Know Your Crew product.

5.2.1 Improving the data-set

The highest priority area of improvement in future iterations of Know Your Crew's Doc2Vec model is the data-set upon which it is trained. Inspection of the results on the query validation sets revealed that the models failed on tests that required it to recognise misspellings or colloquial abbreviations (e.g. 'mgmt' for 'management'). The inclusion of a large Twitter data-set was intended as the remedy for these issues, to capture such common linguistic deviations that are unlikely to be part of

a formally written corpora like Wikipedia. While in this case the inclusion of the Twitter data-set degraded the performance of the model, it is likely that this was due to an ineffective method of parsing and cleaning Tweets. A next iteration of training and validation may try different strategies in parsing and cleaning a Twitter dataset, but may also seek to include another colloquial NLP dataset, for example the Yelp Challenge dataset [52].

It is also considered likely that the of the Doc2Vec model will continue to organically improve (with periodic retraining) as the Know You Crew data-set grows, under the intuition that the greater influence that a data-set has on the neural networks weights the better performance that model will have in associated machine learning tasks on that data-set itself.

5.2.2 Different methods for hyper-parameter optimisation

To avoid a full grid search across the hyper-parameter space, a greedy hill-climbing algorithm approach was employed utilising three runs within randomised initialised parameter settings and order of operation. Although significantly reducing the model training and evaluation time, this process was still incredibly time intensive, and for a next run of the model training process it might be wise to consider other hyper-parameter selection methods. One such possible alternative is randomised parameter optimisation [53], which sets a maximum number of iterations for the parameter search and where each hyper-parameter is randomly selected from a distribution across possible parameter values or from a set of discrete alternatives.

5.2.3 Sentiment analysis model

One of the tasks for which the Doc2Vec model is to be used within Know Your Crews product is to assess the sentiment of responses to certain questions tagged within the Know Your Crew application. Although the results outlined in Section 4.2.4 were encouraging, in that they showed that the Doc2Vec model was significantly

more predictive than the bag-of-words and bi-gram comparators, it did not perform as well as the benchmark of the original Paragraph Vector paper on the benchmark Stanford data-set citele2014distributed (although not directly comparable due to the difference in grain of sentiment classes).

An important point to note is that these are not the same data-sets, and ambiguity in some of the responses to the Know Your Crew questions makes it arguably a more difficult data-set to classify. However, probably the biggest limitation with the Know Your Crew data-set is its size, only just above 400 answers. There is no definitive training set size required for an effective sentiment model, but it is unlikely that this is sufficient. The Stanford data-set for example has over 9,600 sentences. It could be expected that the performance of the model would continue to approach that of the original Paragraph Vector paper as the training set size grows. Rather than relying on the organic growth of the Know Your Crew data-set, one approach considered is to use an online task service, such as Amazon's Mechanical Turk [54] to generate new answers to questions from the Know Your Crew application. One potential function of the sentiment model it to allow managers to view which specific answers generated particular sentiment scores (rather than just a team-wide aggregation) which would require a significantly greater level of confidence in prediction than is currently being achieved on the Know Your Crew data-set.

An additional consideration is that a significantly larger training set would allow for a model per question, rather than a generic sentiment model. This would mean that nuances in the way people response to different questions could be accounted for, or a new style of question could be asked that assessed different kinds of polarity in written language (rather than just positive and negative sentiment).

5.3 Predicting Team Effectiveness

The subsequent task of using in-application metrics to predict self-reported team effectiveness produced one clear stand out result. It found a strong and positive

relationship between the similarity of answers of a team on Personal questions and self-reported team effectivness. Of the other metrics assessed, only similarity on MLT questions produced results of some note, showing a negative correlation with team effectiveness, though not at a statistically significant level.

Although encouraging, there should be caution in interpreting these results as definitive, and employing them in any predictive analytics model. When testing statistical significance across many combination of variables, as the number of combinations increases, so does the chance of encountering a spurious correlation [55].

This exercise has proved useful as a starting point for the company, and consideration is being as to if and how to integrate, and communicate, the finding regarding Personal questions into the Know Your Crew application. However it is also recognised that this study did have some weaknesses, and it is the intention of the company to validate these results in another study which addresses these challenges. Below, some of these issues are discussed, with consideration given to future improved study of team effectiveness.

5.3.1 The type of team

One of the issues identified at the beginning of the study was that it was to be conducted on university teams, rather than workplace teams. This raised several potential problems. The first is that the make-up of a Wharton Business School class is unlikely to reflect the make-up of an average work place team, especially one with a wide range of experience and backgrounds. Secondly, the Wharton teams, by the time of second survey issue, would have only been working together for a few weeks. While this may replicate a Know Your Crew program issued to a new project team, it fails to take account for more established teams. Thirdly, unlike a workplace team, these teams do not work together on a day-to-day basis, and therefore are unlikely to develop a true workplace dynamic.

5.3.2 Self-reporting

This analysis relied on self-reporting of team effectiveness measures by the Wharton cohort - two potential issues were identified relating to this. The first is that as outlined by Hackman (1987) [17] a good measure of team effectiveness should assess the performance of that team from a reliable third party source. An individual within a team in many cases may not be the best to assess the effectiveness of that, particularly if they are the source of ineffectiveness. Secondly, this self-reporting mechanism may have been problematic in that the study contained no definitive mechanism (other than repeated email reminders) to ensure that all participants completed the survey. As was noted in Section 4, only 31 of the original 40 class-size returned the second survey. This was even though rewards were offered for teams that completed both surveys issued during semester.

A next iteration of this study might attempt to access some independent third party data source of team performance. In the Wharton example, this might take the form of end-of-semester student grades

5.3.3 Engagement levels

While there was an initially a high level of engagement with the application in the first few weeks of the study - it was noted that the over time the number of students engaging in gameplay significantly reduced. The Know Your Crew application incorporates mechanisms for encouraging game completion each week, including email reminders and push notifications. For typical workplace teams that use the Know Your Crew application, there is also the pressure from managers to complete the gameplay to ensure that the analytics are correctly reflected. It was perhaps this lack of managerial pressure that could be used to explain the poor engagement of the Wharton Business School students.

Reduced engagement has the effect of limiting the set of questions across which a player could be assessed for any one of the seven independent variabls. This would

likely have increased the variance in measures such as sentiment and similarity, and made it harder to draw correlation between the independent and dependent variables. The nature of this engagement additionally meant it was not possible to draw correlation between responses to specific questions and team effectiveness measures, rather than just aggregate methods utilised.

5.3.4 The sample size

With only 31 students returning the secondary survey, the sample size for analysis was limited. Because of this, there may have been weak, but neverthless signiifcant correlations that were missed as part of this analysis. A second iteration of this study would ideally include a much larger cohort of participants.

5.3.5 Generalisability of results

Questions remain whether the correlation results are generalizble, or whether it was only related to this specific set of questions served to the Wharton students. The results are only truly useful to the company if they generalizable to other questions of this nature, because of the likelihood that over time the content of the Know Your Crew application will evolve.

6. REFLECTIONS & CONCLUSION

6.1 Overview

This section reviews the entire project in terms of the adequacy of the project plan, differences in approach from the original proposal, the extent to which the project's objectives were met, potential future work to stem from the work and the author's personal reflections.

6.2 Reviewing the project plan

Whilst the project plan was adequate, there are two main areas in retrospect that could have been improved. Firstly, more time could have been allocated to the issue of data collection and preparation, particularly considering the difficulty had successfully integrating a colloquial language dataset like Twitter into the analysis.

Secondly, more consideration could have been given to the literature review, specifically referring to the team effectiveness elements. While all the elements assessed had a strong footing in the team effectiveness literature, any future work might benefit from a more top-down and holistic approach to identifying the different elements of team effectiveness that are relevant to the Know Your Crew application.

6.3 Differences in approach from the proposal

The most significant change in this from the original proposal (see Appendix A), was the addition of the survey issued to the Wharton Business School to assess team effectivness. Although such a survey was always an objective of the company, it was initially unknown whether this would be completed within a timeframe suitable for this reports deadlines. Therefore, the initial proposal called for simply assessing the

correlation between metrics such as those produced by the Doc2Vec model, and proxies for team cohesion (just one of seven contributing factors to the team effectiveness measure that was eventually defined for this report). For example, it was thought prediction accuracy may be a proxy of team cohesiveness as it provides some insights into the knowledge team members have of each others preferences and opinions.

The final objectives of this report have also been streamlined from what was included in the initial proposal. The proposal defined four different objectives including a literature review, implementation of a language model, use of that model to complete a document similarity task, and the measurement of team cohesion. As the project was completed it became obvious to think of it in terms of two primary objectives, the training and evaluation of a language model for various machine learning tasks (of which document similarity was one), and the correlative analysis to assess in-application metrics against team effectiveness from the Wharton Survey. The literature review was not ultimately considered an objective but a tool to inform the two primary goals of the analysis.

Another key point of differentiation was the addition of additional independent metrics to undertake a correlative analysis. Initially just the similarity of players responses across all questions was to be considered. However, this was changed to group questions into similar buckets, with the the hypothesis that different question types may derive different types of correlation with team effectiveness. The set of independent variables was also expanded to include sentiment analysis, application engagement and predication accuracy. These decisions were based on inspiration drawn from the body of literature that addresses team effectiveness, as well as hypotheses generated in consultation with the Know Your Crew team.

6.4 Future work

There are a number of elements of future work that can be defined upon completion of this project. The first is the continued improvement and iteration of the

Doc2Vec model as discussed in Section 5. This model is likely to have multiple different applications within the Know Your Crew product. Currently the model has been used to implement and test a sentiment analysis model, as well as a 'similarity' framework to determine how aligned individuals within a team are on certain issues related to both work and non-work issues. There is also currently a plan to use the model's outputs based on Know Your Crew data to be predictive on independent, third-party metrics of personality, to provide managers insights into the distribution of their team on some personality scale.

There is also the intention to utilise the findings from the team effectiveness component of this report specifically relating to the finding that the similarity of a team's response on personal related questions is positively correlated to that team's effectiveness. This work will comprise two components. The first will be to determine the most effective way to validate the results found, using a new study to resolve the issues identified in Section 5 about the original study conducted with the Wharton Business School. The second will be to determine the most useful way to integrate this found knowledge into the Know Your Application whether this relationship is explicitly conveyed to managers or whether it simply forms input to other elements of the company's analytics framework.

6.5 Personal reflections

This project was an interesting undertaking for me personally, and I also believe it has played a very crucial role in my professional development. It has ignited an interest in several different areas, most prominently NLP and Machine Learning. Although many of the processes of this project utilised existing libraries (e.g. Gensim), the understanding I have gained through this piece of work has encouraged me to begin thinking about the ways I could code my own NLP machine learning algorithms, or to begin contributing to the existing open-source libraries.

LIST OF REFERENCES

LIST OF REFERENCES

[1] Google, "Vector representations of words." `https://www.tensorflow.org/versions/r0.10/tutorials/word2vec/index.html`. Accessed on 11/22/2016.

[2] G. Leshed and J. Kaye, "Understanding how bloggers feel: recognizing affect in blog posts," in *CHI'06 extended abstracts on Human factors in computing systems*, pp. 1019–1024, ACM, 2006.

[3] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Citeseer, 2007.

[4] G. E. Hinton, J. L. Mcclelland, and D. E. Rumelhart, "Distributed representations, parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations," 1986.

[5] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[6] J. R. Firth, "{A synopsis of linguistic theory, 1930-1955}," 1957.

[7] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[10] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents.," in *ICML*, vol. 14, pp. 1188–1196, 2014.

[11] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. `http://is.muni.cz/publication/884893/en`.

[12] O. Levy and Y. Goldberg, "Dependency-based word embeddings.," in *ACL (2)*, pp. 302–308, 2014.

[13] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model.," in *Aistats*, vol. 5, pp. 246–252, Citeseer, 2005.

[14] A. Mnih and Y. W. Teh, "A fast and simple algorithm for training neural probabilistic language models," *arXiv preprint arXiv:1206.6426*, 2012.

[15] S. W. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological science in the public interest*, vol. 7, no. 3, pp. 77–124, 2006.

[16] J. J. Phillips, R. Stone, and P. Phillips, *The human resources scorecard*. Routledge, 2012.

[17] J. Hackman, "The design of work teams. inj. w. lorsch (ed.), handbook of organizational behavior (pp. 315-342)," 1987.

[18] V. U. Druskat and S. B. Wolff, "Building the emotional intelligence of groups," *Harvard business review*, vol. 79, no. 3, pp. 80–91, 2001.

[19] F. Luthans, "Organizational behavior: An evidence based approach, 13-th edition," *McGrawHill Irwin. Fq*, vol. 141, 2011.

[20] R. L. Daft, *The leadership experience*. Cengage Learning, 2014.

[21] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe, "Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes," *Journal of Computer-Mediated Communication*, vol. 19, no. 4, pp. 855–870, 2014.

[22] T. Schuller, S. Baron, and J. Field, "Social capital: a review and critique," *Social capital: Critical perspectives*, pp. 1–38, 2000.

[23] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[24] V. U. Druskat, G. Mount, and F. Sala, *Linking emotional intelligence and performance at work: Current research evidence with individuals and groups*. Psychology Press, 2013.

[25] P. Neves and R. Eisenberger, "Perceived organizational support and risk taking," *Journal of Managerial Psychology*, vol. 29, no. 2, pp. 187–205, 2014.

[26] D. H. Lee, "The influence of trust on sharing information,"

[27] L. Prusak, "Building a collaborative enterprise," 2011.

[28] S. Brown, D. Gray, J. McHardy, and K. Taylor, "Employee trust and workplace performance," *Journal of Economic Behavior & Organization*, vol. 116, pp. 361–378, 2015.

[29] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and brain sciences*, vol. 1, no. 04, pp. 515–526, 1978.

[30] C. S. Wang, T. Kenneth, G. Ku, and A. D. Galinsky, "Perspective-taking increases willingness to engage in intergroup contact," *PloS one*, vol. 9, no. 1, p. e85681, 2014.

[31] I. J. Hoever, D. Van Knippenberg, W. P. van Ginkel, and H. G. Barkema, "Fostering team creativity: perspective taking as key to unlocking diversity's potential.," *Journal of Applied Psychology*, vol. 97, no. 5, p. 982, 2012.

[32] S. R. Bray and L. R. Brawley, "Role efficacy, role clarity, and role performance effectiveness," *Small Group Research*, vol. 33, no. 2, pp. 233–253, 2002.

[33] T. Erickson, "The biggest mistake you (probably) make with teams." `https://hbr.org/2012/04/the-biggest-mistake-you-probab`, April 2012. (Accessed on 12/12/2016).

[34] G. Lynn and F. Kalay, "The effect of vision and role clarity on team performance," *Journal of Business Economics and Finance*, vol. 4, no. 3, 2015.

[35] A. Pentland, "The new science of building great teams," *Harvard Business Review*, vol. 90, no. 4, pp. 60–69, 2012.

[36] R. Schwarz, "What the research tells us about team creativity and innovation." `https://hbr.org/2015/12/what-the-research-tells-us-about-team-creativity-and-innovation`, December 2015. (Accessed on 12/17/2016).

[37] L. L. Gilson, J. E. Mathieu, C. E. Shalley, and T. M. Ruddy, "Creativity and standardization: complementary or conflicting drivers of team effectiveness?," *Academy of Management Journal*, vol. 48, no. 3, pp. 521–531, 2005.

[38] S. E. Page, *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press, 2008.

[39] "How diversity makes us smarter - scientific american." `https://www.scientificamerican.com/article/how-diversity-makes-us-smarter/`. (Accessed on 11/22/2016).

[40] R. B. Reiich, "The 'pronoun test' for success - the washington post." `https://www.washingtonpost.com/archive/opinions/1993/07/28/the-pronoun-test-for-success/e45f3343-8b9b-444c-b7c2-2afa235c53e3/?utm_term=.1bc9414ea9cd`, July 1993. (Accessed on 11/22/2016).

[41] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker, "Language style matching as a predictor of social dynamics in small groups," *Communication Research*, 2009.

[42] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, no. 1, pp. 39–44, 2011.

[43] Wikimedia, "Wikimedia downloads." `https://dumps.wikimedia.org/`, 12 2016. (Accessed on 12/11/2016).

[44] S. M. Translation, "1 billion word language model benchmark." `http://www.statmt.org/lm-benchmark/`, 12 2016. (Accessed on 12/11/2016).

[45] V. Minah, *Representation Learning for Structural Music Similarity Measurements*. PhD thesis, City University London, 2015.

[46] "Deeply moving: Deep learning for sentiment analysis," 2016.

[47] E. Salas, R. Grossman, A. M. Hughes, and C. W. Coultas, "Measuring team cohesion observations from the science," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 365–374, 2015.

[48] M. C. Yang and Y. Jin, "An examination of team effectiveness in distributed and co-located engineering teams," *International Journal of Engineering Education*, vol. 24, no. 2, p. 400, 2008.

[49] R. B. d'Agostino, "An omnibus test of normality for moderate and large size samples," *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971.

[50] K. Bowman and L. Shenton, "Omnibus test contours for departures from normality based on b1 and b2," *Biometrika*, vol. 62, no. 2, pp. 243–250, 1975.

[51] Scipy.org, "scipy.stats.mstats.normaltest." `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mstats.normaltest.html`. (Accessed on 12/18/2016).

[52] Yelp, "Yelp dataset challenge." `https://www.yelp.com/dataset_challenge/`. (Accessed on 12/11/2016).

[53] Scikit-learn, "Tuning the hyper-parameters of an estimator." `http://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-search`. Accessed on 12/23/2016.

[54] Amazon, "Mechanical turk." `https://www.mturk.com`, 2016. (Accessed on 12/18/2016).

[55] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLoS Biol*, vol. 13, no. 3, p. e1002106, 2015.

APPENDIX

A. ORIGINAL PROJECT PROPOSAL

THIS PAGE INTENTIONALLY LEFT BLANK

MSc Data Science Project Proposal

Assessing sentence similarity with doc2vec

Edward Atkins
Supervisor: Tillman Weyde

1. Introduction

This project will be completed as part of my internship with Know Your Crew (KYC). KYC provide a gamified mobile platform that helps team connect and build trust, and give managers actionable analytics to optimize team dynamics.

The basic functionality of the game platform is that players respond to questions generated by the application (e.g. 'What are three things that our team could do better?') with a free text response. Every play must then guess which other player provided each answer.

The free text nature of this response provides some interesting opportunities, but also some significant challenges in analyzing the data generated by the platform. One of the core challenges for KYC is how to utilize the textual data for different data analysis and machine learning tasks. One of these areas of interest for KYC is whether socially connected players can be identified through similarity of their answers to a selected set of questions. Ultimately it is the intention of the company to use this information to inform team creation and management.

Determining the semantic similarity of natural language is not a straight forward task. For example, the phrases 'gone jogging' and 'having a run' mean very similar things, while not sharing a single common word. To tackle this problem and determine the similarity of answers within the KYC platform, I propose to implement the doc2vec algorithm, an extension of the popular word2vec algorithm. The model will output a vector representation of a given answer, which allows answers to be compared for their similarity.

The similarity of users answers across multiple questions, will be tested for correlation with existing metrics that are thought to indicate a relationship within the application, for example bi-directional accuracy prediction.

As part of my internship for KYC I will also be likely testing the applicability of the doc2vec model as a feature extraction method for sentiment analysis tasks, but this will not form a part of my thesis, except as a tool to describe model accuracy.

2. Objectives and purpose

Using the categories outlined by Oates (2006, p.16-21), I define the primary purposes of my research paper as being to 'add to the body of research', 'to solve a problem' and to 'contribute to personal needs'.

Add to the body of research
I believe my project will add to the body of research in two main ways. Firstly, it presents an application of the doc2vec algorithm to a unique dataset. Secondly, I will explore methods of

assessing the accuracy of the doc2vec algorithm, which are not currently extensive addressed in the current body of research.

Solve a problem
The project will directly seek to address the problem of how to assess the similarity of KYC users' answers free text answers.

Contribute to personal needs
Personally, this research covers two areas of my own interest, Natural Language Processing (NLP) and Artificial Neural Networks (ANN), and I believe it will demonstrate to potential employees my interest and competency in these two domains.

Below I define the primary objectives of my research as well as the criteria against which success will be measured for each:

	Research Objective	**Outcome & Measure**
1	Complete a review of existing applications and evaluation techniques of both the word2vec and doc2vec algorithms	The output will be a document that comprehensively describes the state of research in the domain
2	Implement doc2vec algorithm using existing frameworks, training with combination of external and KYC data	The output will be a doc2vec model saved to disk, capable of being utilized for various machine learning tasks within the company. Success will be measured by testing its word vector properties on a dataset of semantic questions, and its document vector properties in a basic sentiment analysis task. Both of these can be compared to existing benchmarks.
3	Employ model to assess similarity of player's answers in KYC application	The output will be an implementation that allows comparison of any two answers from the KYC application for similarity. The efficacy of the model in this task will be assessed using a custom built test data set using real answers from the KYC application.
4	Determine relationship between the answer similarity of teammates and existing social metrics	The output will be a statistical analysis of the relationship between answer similarity and other in-app metrics. Success will be the identification of metrics that are highly correlated with answer similarity.

The beneficiaries of this research will primarily be KYC.

3. Critical Context

Traditionally Natural Language Processing (NLP) systems treat different words as 'discrete atomic symbols' (Tensorflow.org, 2016), with no representation of the implicit relationships that exist between words. For example, the bag-of-words model utilizes a count of every word in the known vocabulary to undertake machine learning tasks.

Word2vec

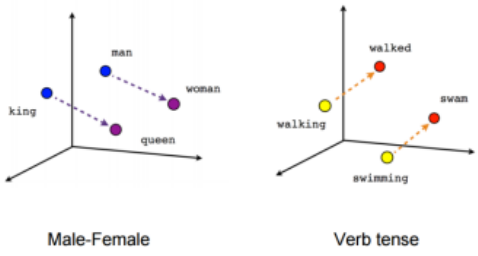
A more sophisticated technique that has come to prominence in recent years is the word2vec model (Mikolov et al., 2013) that learns vector representations of words, also known as 'word embeddings'. Rather than simply representing words as a binary switch word count, the model learns and implicitly represents the relationships between words, for example that 'Paris' and 'Rome' are both words that represent Capital Cities, located in Europe.

The word2vec model is an example of a Vector space models (VSMs) 'represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points' (Tensorflow.org, 2016). Using the above example, it would be expected that while 'Paris', 'Rome' and 'Sydney' would all be mapped relatively close together because they are all capital cities, Paris' and 'Rome' would be some degree closer because they are more similar (i.e. both located in Europe).

Word2vec actually refers to two different algorithms, a Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. CBOW predicts target words from its surrounding words, while Skip-Gram does the obvious. Both models are shallow neural networks, with one hidden layer, that are trained using a binary classification objective to identify real target words from imaginary words.

This process trains a hidden layer, its vector representation, for each word. When these vectors are visualized in 2 dimensions that they capture some intuitively useful semantic relationships, as is shown in Figure 1 (Tensorflow.org, 2016).

Figure 1 – Semantic Relationships



Doc2vec

While the relationships demonstrated above are interesting, for machine learning tasks the natural language object in question is typically a sentence, paragraph or document rather than a single word. To transform multiple word vectors composing a longer document into a single feature vector researching have tried different methods including weighted averages of all words in a document or combinations by the order given by the parse tree of a sentence (The Morning Paper, 2016). However, for classification tasks these methods have proved no more effective than a bag-of-words model (Le and Mikolov, 2014).

To address this issue Le and Mikolov (2014) developed the 'Paragraph Vector' (or 'doc2vec') model an extension of the word2vec distributed word model. To predict the next word in a sentence, the model now relies on a combination (concatenation or averaging) of *both* a word vector and a paragraph vector.

Model evaluation
There are several methods that have been employed for evaluating the effectiveness of the word2vec and doc2vec models.

Perhaps the simplest is to assess their performance in classifying the sentiment of a benchmark textual dataset. Le and Mikolov (2014) for example show that the doc2vec model significantly outperforms bag-of-words and bag-of-bigrams methods in classifying a dataset of IMDB movie reviews. As discussed previously, the word2vec model (using a weighted average method) shows no improvement over these other techniques.

To assess the accuracy of the word2vec model, a common methodology is to test its relational properties, by asking semantic questions (Mikolov et al., 2013). For example, the question, 'man is to king as is to queen?' can asked as:

X = Vector('king') – Vector('man') + Vector('queen')

If the algorithm has been trained on a suitable dataset, the closest vector in the model's vocabulary to the newly calculated X should be 'woman'. Questions such as this, where there is a definitive 'best' answer can be asked of the model and an accuracy percentage calculated.

There is no such property that can be tested directly with the doc2vec algorithm. Instead it is common for its accuracy to be calculated through an information retrieval task. Le and Mikolov (2014) use a dataset of paragraphs returned from search engine queries. For each query they generate two paragraphs of that query and a randomly selected query, with the task being to determine which paragraphs are from the same query.

4. Approaches: Methods & Tools for Analysis & Evaluation

Below I have outlined the key components of my methodology, roughly in line with the Plan of Work described in Section 5. Some key methodological choices will not be finalized until the first weeks of the project, but where possible I indicate my initial thinking.

Literature Search & Review
I have already undertaken a literature review to provide critical context for this proposal and to develop my methodology. This will be built out in the first stages of my project timeline, with a focus on implementations of the doc2vec algorithm as well as methods to assess the accuracy of this algorithm.

Data sources and preprocessing
The word2vec and doc2vec algorithms typically are trained on datasets of several billion words (Kim, 2014). Therefore, the KYC dataset on its own will not be sufficient in its own right. To train the model I will utilize not only the KYC dataset, but also other publicly available datasets. These will likely include Google News, Wikipedia, (Code.google.com, 2016) as well as other sources, like the Twitter API that may be more reflective of the text samples that are generated by KYC users.

The chosen data sources are likely to be diverse, and will each require a degree of preprocessing to conform them to a format suitable for input into the machine learning algorithm.

Implementation of models

It would be possible to develop and train my own word2vec or doc2vec model using a deep learning framework such as Theano or TensorFlow. However, there would be no practical advantage to this, as there are already existing frameworks that implementation of the algorithms.

One of the major frameworks that provides an implementation of both the word2vec and doc2vec algorithms is Gensim (Řehůřek and Sojka, 2010). Gensim's natural language processing language is optimized by employing Cython, which provides results 24 times faster than baseline Numpy code.

Gensim also provides a simple API to implement the algorithm using a distributed computing framework such as Amazon Web Services (AWS). Depending on the local performance in tests of the training phase of the algorithm, a service such as AWS may be utilized.

Model evaluation

To evaluate the performance of the doc2vec model I propose to draw from the information retrieval technique used by Le and Mikolov (2014), but make it more specific to my dataset and task.

I will use real answers generated from the KYC application to develop a number of different groupings of answers and use these as the basis of my test set. For example for a question relating to a user's preferred morning routine, I my create groups relating to breakfast (e.g. 'coffee and toast', 'eat bacon and eggs' etc.), exercise (e.g. 'go for a morning jog' , 'get up and go to the gym' etc.) and work related (e.g. 'check my emails', 'review my calendar, check emails' etc.). I will then generate combinations of the form:

Test Response: Get up and go to the gym
Similar Response: Go for a morning jog
Non Similar Response 1: Check my emails
Non Similar Response 2: Eat bacon and eggs

The algorithm will be used to judge the similarity of the test response to the three other responses in terms of the distance of their generated document vectors and the accuracy across a large number of questions and combinations of answers recorded.

This custom evaluation methodology will be designed to mimic as closely as possible its real-life application of judging the semantic similarity of two users answers to a specific question. I will perform a cross-validation process across different hyper-parameters (e.g. vector size and min word-count) as well as assessing the effectiveness of different combinations of the datasets I propose to use.

The performance of the final model will be compared against other basic NLP methods e.g. bag of words, and bag of n-gram methods.

KYC intends that the trained doc2vec model be also utilized for a sentiment analysis task. Although this is not part of the scope of my thesis, I believe it will provide an effective illustration of the efficacy of the algorithm, against well-known benchmarks.

Assessing social connections

Finally, I will be using the similarity of users answers to assess the strength of their social connection.

The ultimate intention of KYC is to issue a survey pilot to answers, to provide a ground truth upon which a machine learning model can be built to predict the strength of these connections. However, the timing is indefinite, and so for the moment it is not included in my thesis scope, although it may later be included.

For the moment it is my intention to conduct a basic analysis that compares the similarity of users answers across a set of questions with other existing metrics within the KYC application that could be considered to be related to social connectedness. For example, whether they are member of the same team, accuracy of prediction with each other, in-game interactions such as chat and 'nudges'.

To determine their relationship, I will use basic statistical tools such as Pearson's correlation coefficient to assess the strength of the relation.

Ethical, Legal & Professional Issues

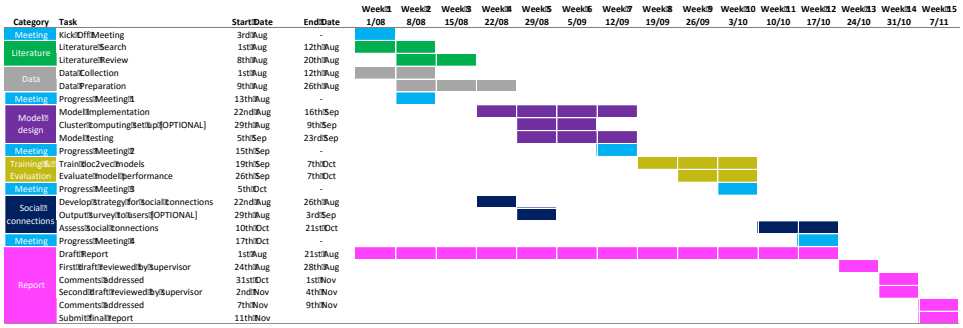
In developing this methodology, consideration was given to City University's Research Governance Framework. Importantly this project does not include any human participants.

After completing the University's Ethics Review Form (see Appendix) and considering other areas of potential legal and professional significance, no issues were identified for this research.

5. Plan of Work

The Gantt chart below outlines my proposed work plan for successful completion of this project. The project will start on 1st August 2016 and I plan to have submitted my final report by the 11th November 2016. This leaves one week as contingency for any unforeseen issues. The final due date for submission is the 6th January 2016, so I am leaving myself flexibility if there are changes to the project scope through my work with KYC.

As well as a kick-off meeting in the first week of the project, I will have four subsequent progress meetings with my supervisor over the length of the project.



Category	Task	Start Date	End Date
Meeting	Kick Off Meeting	3rd Aug	-
Literature	Literature Search	1st Aug	12th Aug
	Literature Review	8th Aug	20th Aug
Data	Data Collection	1st Aug	12th Aug
	Data Preparation	9th Aug	26th Aug
Meeting	Progress Meeting 1	13th Aug	-
Model design	Model Implementation	22nd Aug	16th Sep
	Cluster computing Set Up [OPTIONAL]	29th Aug	9th Sep
	Model testing	5th Sep	23rd Sep
Meeting	Progress Meeting 2	15th Sep	-
Training & Evaluation	Train Doc2vec models	19th Sep	7th Oct
	Evaluate model performance	26th Sep	7th Oct
Meeting	Progress Meeting 3	5th Oct	-
Social connections	Develop Strategy for social connections	22nd Aug	26th Aug
	Output survey to users [OPTIONAL]	29th Aug	3rd Sep
	Assess social connections	10th Oct	21st Oct
Meeting	Progress Meeting 4	17th Oct	-
Report	Draft Report	1st Aug	21st Aug
	First Draft reviewed by supervisor	24th Aug	28th Aug
	Comments addressed	31st Oct	1st Nov
	Second Draft reviewed by supervisor	2nd Nov	4th Nov
	Comments addressed	7th Nov	9th Nov
	Submit final Report	11th Nov	

6. Risks

Using Dawson's (2006) framework for risk management, I have identified a set of potential risks for my project. I have outlined the nature of the risk, its risk likelihood, risk consequence, risk impact (likelihood * consequence). I describe the mitigating steps I will take to reduce either the likelihood or consequence of these risks as well as as any contingency plan if the risk eventuates.

#	Risk Description	Likelihood	Consequence	Impact	Mitigation	Contingency
1	**Disruption due to illness or personal reasons**	**2**	**4**	**8**	Significant contingency period between planned project end date	Scale down the project
2	**Incorrect estimation of time for task**	**4**	**2**	**8**	Careful consideration has been given to the project plan, with given times on the high side of my initial estimates	Tasks with greatest uncertainty built in to first half of the project, so project timeline may be re-evaluated early in the project
3	**Implementation of genism implementation proves too hard**	**1**	**4**	**4**	I have collected a large number of materials relating to different implementations of this model to assist me	KYC have offered to provide me assistance from their different industry connections
4	**Algorithm takes too long to train**	**4**	**3**	**12**	If required, utilize Amazon Web Services to run algorithm on EC2 instance	Reduce size of training set
5	**Hardware failure or accidental deletion of work**	**2**	**4**	**8**	Use of version control software (git/GitHub) and periodic manual backup	Restore from backup

References

Code.google.com. (2016). *Google Code Archive - word2vec*. [online] Available at: https://code.google.com/archive/p/word2vec/ [Accessed 25 Jul. 2016].

Kim, Y. (2014). Convolutional neural networks for sentence classification. In: *EMNLP*. pp.1746–1751.

Le, Q. and Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*. ICML.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*. [online] Available at: http://arxiv.org/pdf/1301.3781.pdf [Accessed 20 Jul. 2016].

Mikolov, T., Yih, W. and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In: *Conference of the North American Chapter of the Association for Computational Linguistics*.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In:*Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. [online] ELRA, pp.45-50. Available at: https://radimrehurek.com [Accessed 25 Jul. 2016].

Tensorflow.org. (2016). *Vector Representations of Words*. [online] Available at: https://www.tensorflow.org/versions/r0.9/tutorials/word2vec/index.html [Accessed 20 Jul. 2016].

the morning paper. (2016). *Distributed representations of sentences and documents*. [online] Available at: https://blog.acolyer.org/2016/06/01/distributed-representations-of-sentences-and-documents/ [Accessed 25 Jul. 2016].

Appendix: Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

Part A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

Part B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in this case. The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses.

A.1 If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval.		
1.	Does your project require approval from the National Research Ethics Service (NRES)? For example, because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/.	**No**
2.	Does your project involve participants who are covered by the Mental Capacity Act? If so, you will need approval from an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/.	**No**
3.	Does your project involve participants who are currently under the auspices of the Criminal Justice System? For example, but not limited to, people on remand, prisoners and those on probation? If so, you will need approval from the ethics approval system of the National Offender Management Service.	**No**

A.2 If your answer to any of the following questions (4 – 11) is YES, you must apply to the City University Senate Research Ethics Committee (SREC) for approval (unless you are applying to an external ethics committee).		
4.	Does your project involve participants who are unable to give informed consent? For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?	**No**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**

6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**
7.	Does your project involve participants disclosing information about sensitive subjects? For example, but not limited to, health status, sexual behaviour, political behaviour, domestic violence.	**No**
8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (See http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve physically invasive or intrusive procedures? For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**No**
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**

A.3 If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**No**
14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a specific course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**

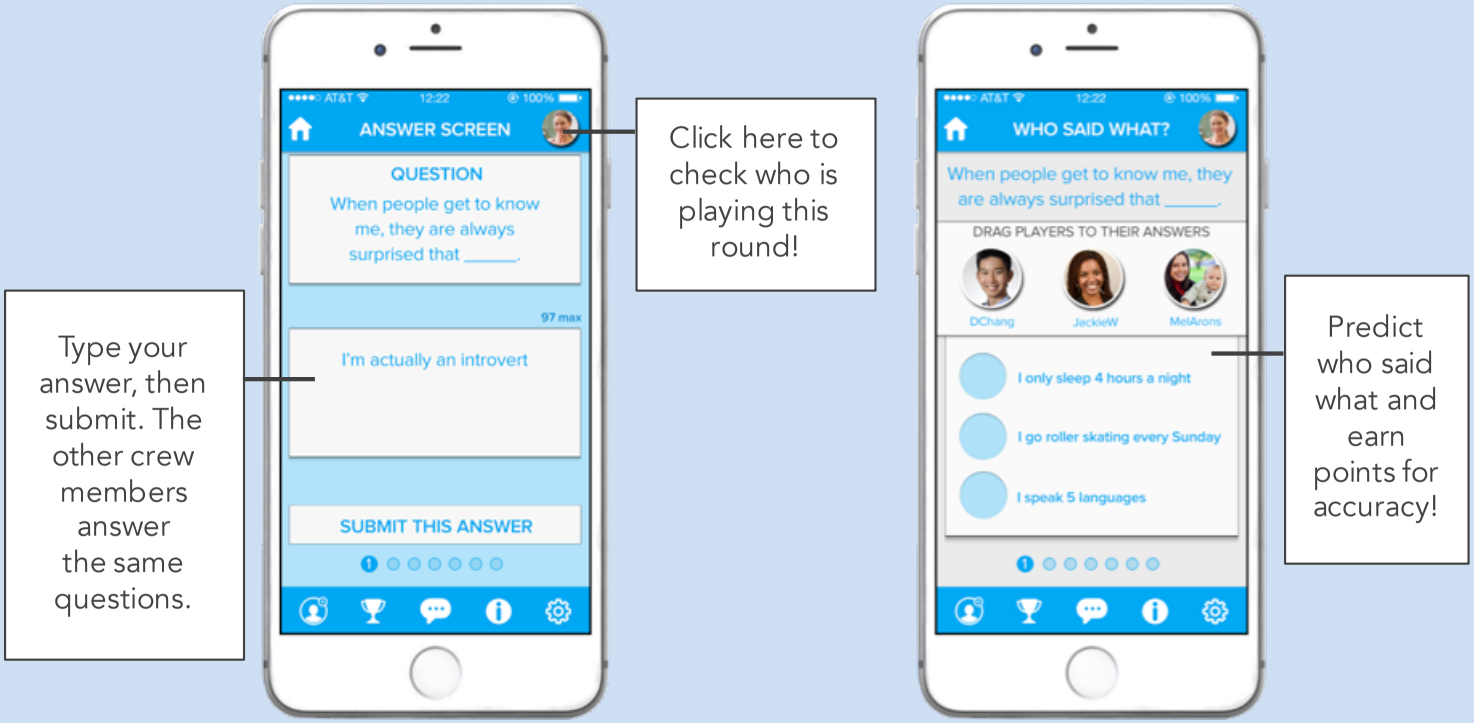
A.4 If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants or their identifiable personal data? For example, as interviewees, respondents to a survey or participants in testing.	**No**

B. KNOW YOUR CREW GAME-PLAY

THIS PAGE INTENTIONALLY LEFT BLANK

How To Play

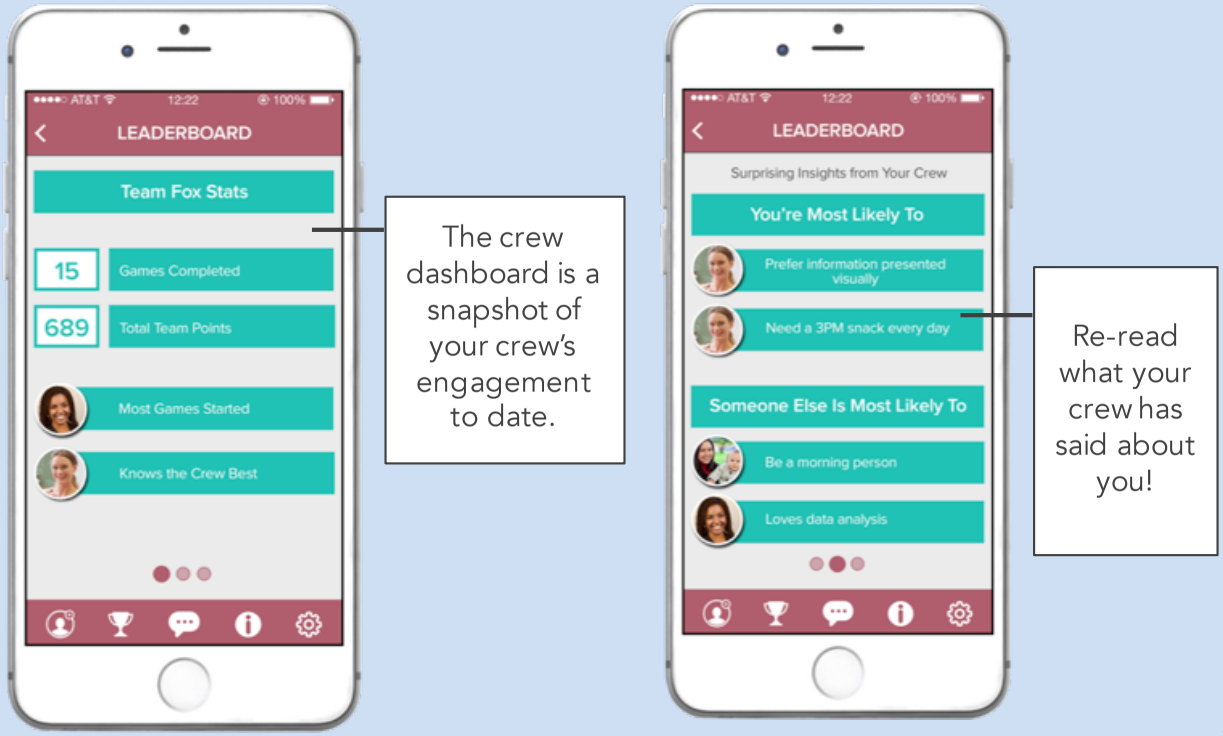
Every game is **5 questions** and users receive **1 game** per week. Once everyone answers the questions, users return to the game to predict who said what – for points!



14

Insights: App Leaderboard

Users earn points based on prediction accuracy and engagement.
Users head to the leaderboard for crew totals and personal insights.



C. LETTER FROM KNOW YOUR CREW

THIS PAGE INTENTIONALLY LEFT BLANK

January 5, 2017

To whom it may concern,

I can confirm that Edward undertook the work detailed in this thesis.

While we have permitted him to submit his code-base as an attachment to his report; we were unable to release any of the accompanying data, due to its confidential nature.

If you have any questions, please feel free to contact me on +1 (202) 341-5551.

Sincerely,

Alison Bloom-Feshbach
CEO, Know Your Crew Inc.
alison@knowyourcrew.com