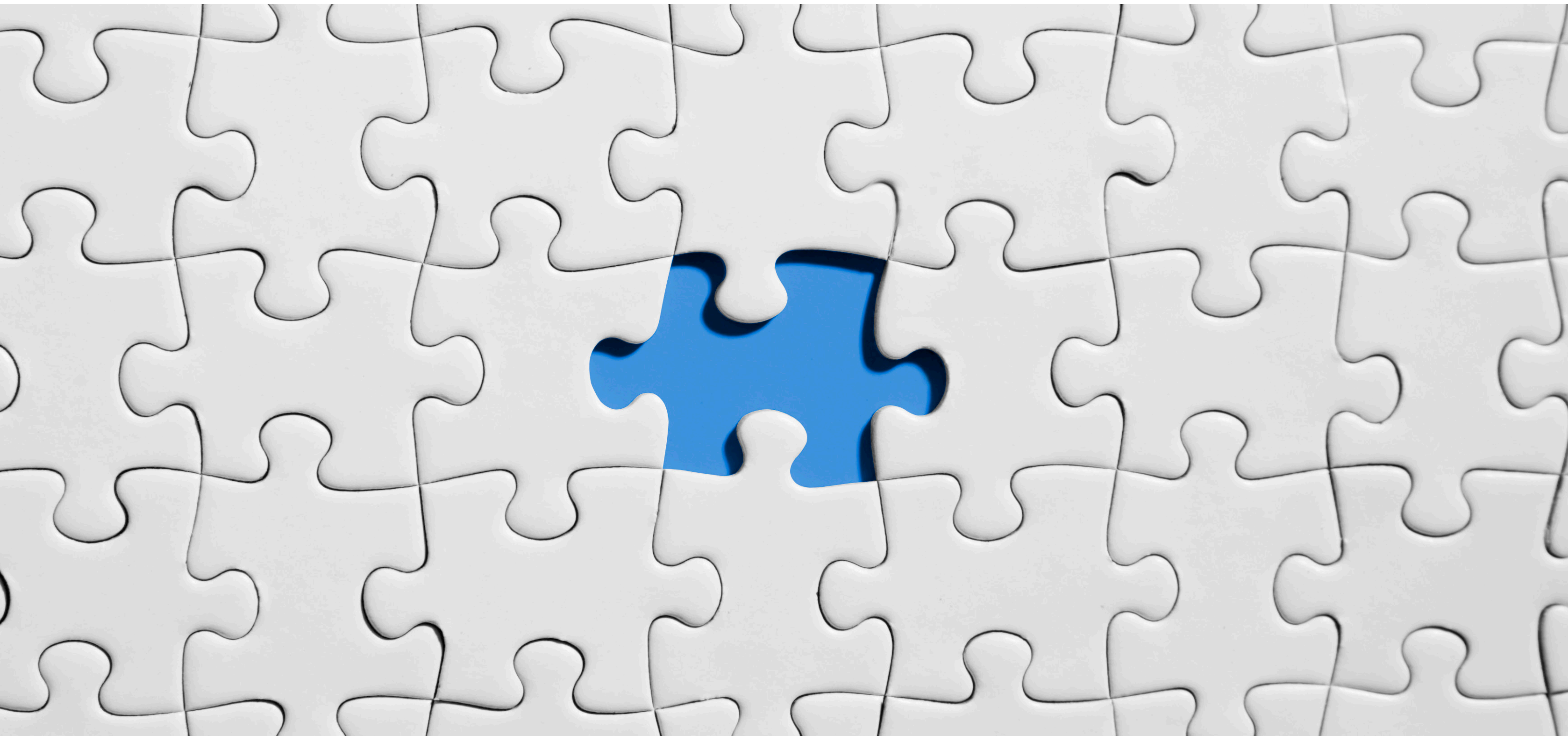
CUSTOMER PROSPECTING FOR A BIOTECH COMPANY



Tony Ku

Advisory Consultant

Dell EMC

tony.ku@dell.com

Fernanda Campello DeSouza

Senior Consultant

Dell EMC

fernanda.campellodesouza@dell.com

Marie Breton

Senior Solutions Principal

Dell EMC

marie.breton@dell.com

Wei Lin

Senior Manager, Chief Data Scientist

Dell EMC

w.lin@dell.com

Knowledge Sharing Article

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell EMC's views, processes or methodologies.

1. Introduction

Medical research is heavily funded around the world. In the United States, the National Institutes of Health (NIH) invests around $32.3 billion per year in medical research funding, reaching more than 2,500 research institutions (National Institutes of Health (NIH) , 2016), a large part of which are government and university research institutes. Many research projects in this field rely on the use of reagents and test kits for laboratory experiments, which are manufactured and sold by biotechnology companies. Currently the biotech company relies on its representatives and their close relationships with researchers to educate the researches and generate product demand that in turn drive product sales. This method is time consuming and expensive for the biotech company to maintain. Using data analytics to identify potential customers and predict product needs could lead to reducing the cost of sales while increasing sales.

One of the first challenges a biotech company must overcome is to identify key influencers in the research community. A second challenge is to identify the potential users of their products. Point of Sale (PoS) systems and customer relationship management (CRM) systems often do not provide full visibility to this as most of the orders would be placed by contract officers, purchasing agents, lab managers or postdoctoral fellows. As a consequence, CRM systems will often not give the biotech company full visibility into how its products are being used, and by whom. But since publishing scientific contributions and grant award information is a crucial part of most government and academic institution mandates, there is a large amount of publicly available information that can be leveraged to gain insight to products, end users and applications.

In this paper we demonstrate how publically available information can be used by biotech companies for prospecting customers. Our framework is as follows: (1) a methodology to identify biotech products used in a paper; (2) a methodology to identify researchers who are key influencers in product-specific research networks; (3) a methodology to identify new product-specific keywords used in papers; (4) a methodology that uses text mining to indicate how closely a grant abstract relates to specific biotech products. Our methodology is summarized in Figure **1**.

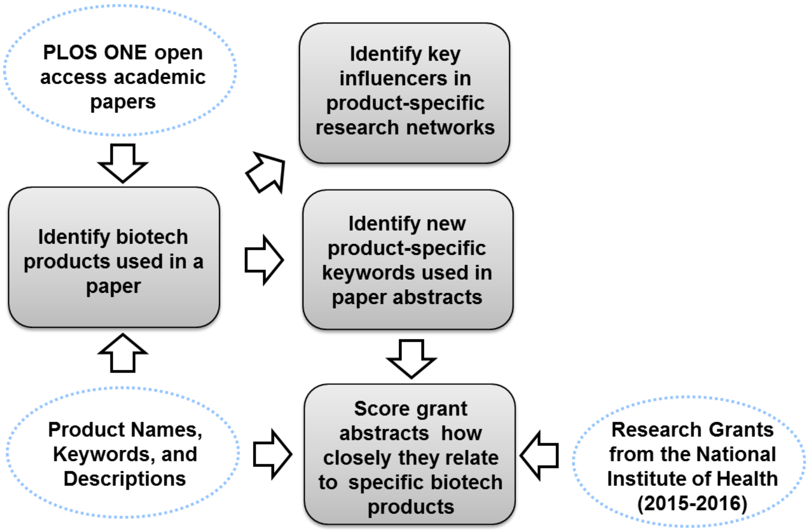

PLOS ONE open access academic papers
Identify biotech products used in a paper
Product Names, Keywords, and Descriptions
Identify key influencers in product-specific research networks
Identify new product-specific keywords used in paper abstracts
Score grant abstracts how closely they relate to specific biotech products
Research Grants from the National Institute of Health (2015-2016)

Figure 1 Methodology summary.

2. Data Sources

The data sources for this paper are:

- **Grant award information from the National Institutes of Health (NIH) in 2015 and 2016:** downloaded in CSV format from the NIH ExPORTER Data Catalog (National Institutes of Health (NIH))
- **Full-text scientific papers from the PLOS ONE open-access portal (http://journals.plos.org/plosone):** downloaded using paper searching and retrieving functions from the R package rplos (Scott Chamberlain, 2016)
- **Scientific paper author information from PubMed in 2015 and 2016 (https://www.ncbi.nlm.nih.gov/pubmed):** downloaded using paper searching and metadata retrieving functions from the R package RISmed (Kovalchik, 2016)
- **Product catalog information from a biotech company website**

3. Biotech product identification

It is common practice for papers in the life sciences to directly reference the specific lab materials (test kits, reagents, etc.) used in the experiments performed, including the manufacturer name. That allows us to identify papers that use products from a particular biotech company through a word search of the company name in the paper's "Materials and Methods" section. Having this scored data set is important in the next step where this information is used as input into the model. For the present illustration we collected all 4,063 papers from the open-access portal PLOS ONE that were published in 2015-2016 and cited a specific biotech company, which we'll refer to as "BioCo", in the "Materials and Methods". We then proceeded to associate the papers to 1,230 specific BioCo products usage as detailed below. The same methodology can be used to examine other publication sources.

We automatically associated papers to specific BioCo products usage using the following steps

1) Create a lexicon of all terms present in BioCo product names, keywords, and descriptions, that are also present in at least one PLOS ONE paper ($N = 2{,}244$ terms).
2) Represent each of the 1,230 products as a binary *N*-dimensional vector indicating whether each term in the lexicon is present in that product's name, keywords, or description $\bar{X} = (x_1 \dots x_N)$.
3) Represent each of the 4,063 papers as a binary *N*-dimensional vector indicating whether each term in the lexicon is present in the set of "Materials and Methods" sentences where the word "BioCo" appears $\bar{Y} = (y_1 \dots y_N)$
4) Compute the distances between product-vectors from Item 2 and paper-vectors from Item 3 using the cosine measure, which measures the angle between documents (Aggarwal, 2015):

$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{N} x_i \times y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

5) Associate each paper to the three products with lowest distance values to that paper

Note that choosing to associate only the three products with smallest distances to a paper is an approximation, since different papers may use different numbers of products from BioCo. For papers using more than three products from BioCo, some valid paper-product usage associations are excluded, whereas for papers using less than three products from BioCo some invalid product-paper usage associations are included. But given that the association is based on terminology, the invalid paper-product usage association may still be a valid paper-product association in terms of relevant subject area association, i.e. that paper may still bring relevant information about the community using that type of product. An alternative methodology to derive product-paper associations would be based on supervised learning applied to a training set of papers with products manually associated by a subject matter expert (which is not currently available).

4. Key influencers identification

We use the product-paper association developed in Section 3 to examine collaboration networks of co-authors associated with specific BioCo products as follows:

1) Compile the list of first authors of all papers associated to that BioCo product in Section 3
2) Select only first authors that were affiliated to institutions in the United States at time of publication. This step serves to narrow the focus to the United States market, as well as to mitigate problems arriving from multiple identically named authors
3) Collect PubMed metadata on all papers published in 2015-2016 and authored by authors selected in Item 2.
4) Using the metadata collected in Item 3, create a co-authorship collaboration network related to the product where each node represents an author and an edge exists between two authors if they co-authored at least one paper. Figure 2 shows an example of collaboration network for Product A.

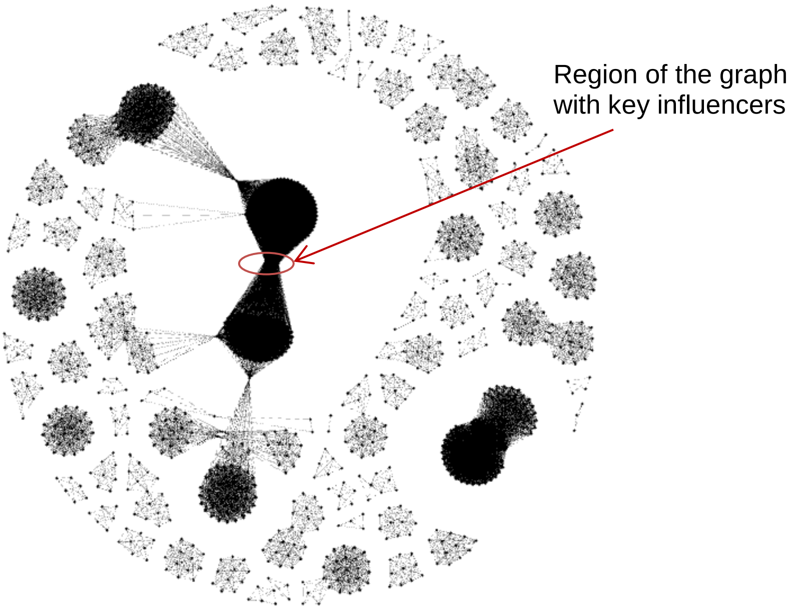


Figure 2 Collaboration co-authorship network for BioCo Product A.

We can examine the collaboration networks to identify key influencers that could be good targets for marketing campaigns and for becoming brand ambassadors. One measure that can be used for that is betweeness centrality, which relates to the number of shortest paths between nodes in the network that pass through a specific node (Aggarwal, 2015). Nodes with high betweenness centrality have high control over the flow of information in the network. Figure highlights the region with nodes of high betweeness centrality.

5. Score NIH grant abstracts for matches to product keywords

The goal of this analysis is to provide the sales team with focused and timely market potential to pursue by scoring grant abstracts on how they match to product keywords. Previously, to help the sales force target their products for areas of research, the company's product management team has defined a set of keywords and general description for each product, which is the base for our first approach of matching to NIH grant abstracts.

5.1. Model

We use about 58,000 NIH grant abstracts (National Institutes of Health (NIH)) in fiscal year 2015 and 2016 as the corpus to perform text mining and scoring, excluding any abstract with fewer than 50 words as it is too short to provide any meaningful content for a research and most likely it's a single line item for personnel hiring (e.g. postdoc for 6 months).

For each term in an abstract, its term frequency (TF) within an abstract and inverse document frequency (IDF) across all abstracts within this corpus are calculated. The product of a term's term frequency and inverse document frequency is the term's term frequency-inverse document frequency (TFIDF). TFIDF is very simple but robust texting mining technique to determine the importance of a term. We use a term's TFIDF as the weight for score calculation for matched key words.

There are several variations of term frequency and inverse document frequency used depending on the circumstances (Wikipedia). Denoting a term as t, a document as d, and the corpus as D, below is the definition we use in this paper.

Term frequency: $TF(t,d) = \frac{t_d}{|d|}$

where t_d is the number of times that term t appears in document d, and $|d|$ is the number of terms in d.

Document Frequency: $DF(t,D) = \frac{t_D}{|D|}$,

Where t_D is the number of documents have the term t at least once, and |D| is the number of documents in D.

Inverse Document Frequency: $IDF(t,D) = \log(\frac{|D|+1}{DF(t,D)+1})$

Note that a smoothing term is used to avoid dividing by zero for terms outside the corpus, for example, new terms appearing in new abstracts. Finally, term frequency-inverse document frequency is defined as:

$$TFIDF(t,d,D) = TF(t,d) * IDF(t,D)$$

To summarize, here are the steps of preprocessing for NIH grant abstracts:

- Sample 58,000 NIH abstracts with 50 or more words, from fiscal year 2015-2016 as the corpus
- Get rid of punctuations but keep hyphens for compound words that are very common in the scientific world
- Remove 153 common stop words in English
- Tokenize into bag of words (unigram)
- Calculate TF, IDF, and TF-IDF for each abstract/term across the entire corpus

Denoting a product as p, an abstract to product score, S, is defined as:

$$S(p,d,D) = \sum\sum TFIDF(t_i,d,D) * \ \delta_{ij}\,(t_j,p), \text{ where } \delta_{ij} = \begin{cases} 1\ if\ t_i = \ t_j \\ \quad 0\ else \end{cases}$$

Because biotech research is the company’s focus, we further filter relevant NIH grant abstracts for scoring to 34,000 based on funding group and project type (IC Name and Activity Type in NIH database).

5.2. Scoring results using company-defined keywords

Figure 3 consists of box plots for five representative product scores against all 34,000 abstracts using company-defined keywords, from very good (Cell Signaling) to fair (Cell Viability Assays) to very poor (Polymerase Chain Reaction). Inter-product scores variability could be due to: 1) different levels of interest across products; or 2) different keyword effectiveness across products. Opportunities for keyword enrichment either manually via crowdsourcing to subject matter experts (SME’s), or automatically via text mining of relevant publications (more to come on this).

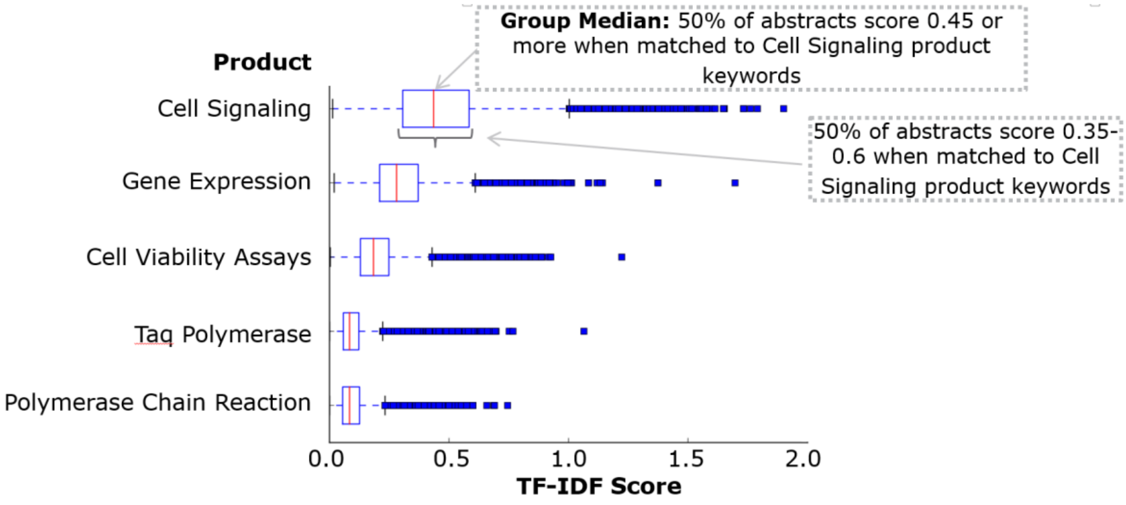
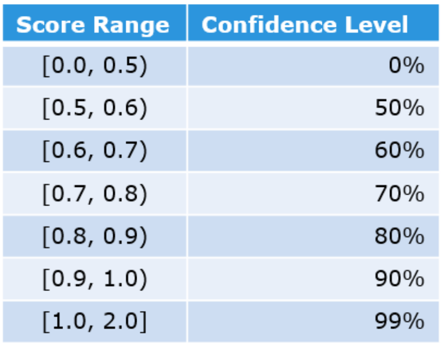


Figure 3 TF-IDF score distribution by product group.

To evaluate the model's effectiveness and determine score threshold per product, we have randomly sampled 100 abstracts out of the top 20% matching percentile for Cell Signaling product with the following stratification (score x 10, sampling %): [(8, .005), (9, .025), (10, .05), (11, .1), (12, .2), (13, .3), (14, .5), (15, .6), (16, .7), (17, .8), (18, 0.99)]. SME's manually review those abstracts to provide match evaluation although this is subjective but the best we can have.

Based on SME's evaluation, the following confidence level has been constructed:



Score Range	Confidence Level
[0.0, 0.5)	0%
[0.5, 0.6)	50%
[0.6, 0.7)	60%
[0.7, 0.8)	70%
[0.8, 0.9)	80%
[0.9, 1.0)	90%
[1.0, 2.0]	99%

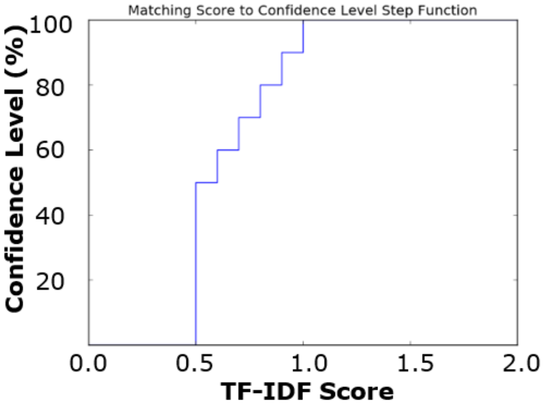
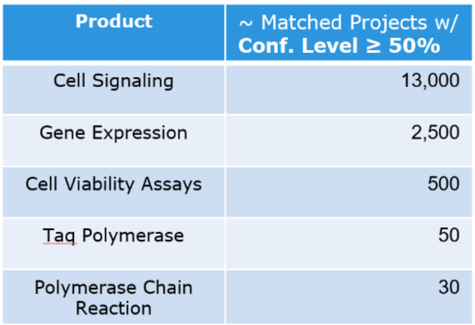


Figure 4 Confidence level construction.

With the confidence level above, we are able to determine the number of matched abstracts per product below:



Product	~ Matched Projects w/ Conf. Level ≥ 50%
Cell Signaling	13,000
Gene Expression	2,500
Cell Viability Assays	500
Taq Polymerase	50
Polymerase Chain Reaction	30

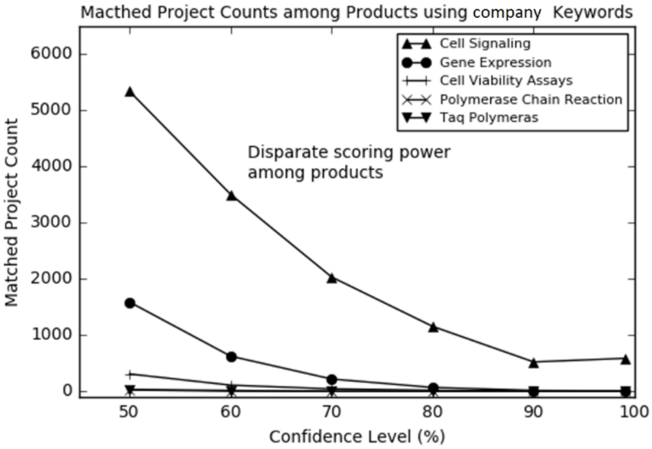


Figure 5 Project matching count.

Cell Signaling product is matched to the largest number of NIH projects as its keywords are much more robust than those of other product groups. The disparate of scoring power is very significant (several orders of magnitude) among products. Keyword enrichment will be discussed in a later use case. It is possible to match multiple products to a NIH project with various confidence levels.

This model could be used to score newly awarded NIH projects as published on a weekly basis to provide the sales team with timely and focused market intelligence. The sales team would then have the opportunity to market their products post-award and before the purchasing agent purchased the necessary products.

5.3. Scoring results using harvested publication keywords

As shown in the previous section, there are several drawbacks using predefined product keywords for matching abstracts:

- Indirectness: NIH grant abstracts are a high-level description of research projects for applying grants, whereas, product keywords are lower-level terms for biotech techniques usually in a publication's method and material section. Thus, either there are very few direct links between these grant abstracts and product keywords or inferencing power is very inconsistent, that is, varies from product to product.
- Staleness: once the task of defining product keywords is completed, they represent a snapshot and tend to stay the same for years to come.
- Limited in scope: because product keywords are defined by several internal SMEs, they are limited to their knowledge. In addition, the company has limited visibility as new applications or protocols have been implemented using their products.

Thus, we employ a different and more robust approach for harvesting product keywords through publications in this section.

As described in Section 3, we are able to identify products cited in a publication's methods and materials section. In turn, more robust and relevant keywords can be harvested from these publications' abstract continuously. Usually one publication can be associated with multiple products and competitor's products.

Here are the steps to harvest product keywords from PlosOne publications:

- Sample 18,000 PlosOne abstracts from year 2015-2016 as the corpus. There are about 4,000 publications associated with 3 different products.
- Get rid of punctuations but keep hyphens for compound words which are common in the scientific world
- Remove 153 common stop words in English
- Tokenize into bag of words (unigram)
- Calculate TF, IDF, and TF-IDF for each abstract/term across the corpus
- Discard terms that appear less than 5 times in the corpus to exclude extremely rare terms
- Calculate TFIDF score per product and term
- Select the terms that are more relevant (with score greater than 0.4) as the harvest keywords for each product

Below is the comparison of scoring power between keywords defined by the company's SMEs and those harvested from PlosOne publication in 3 products which were scored from well, fairly, to poorly using company defined keywords:

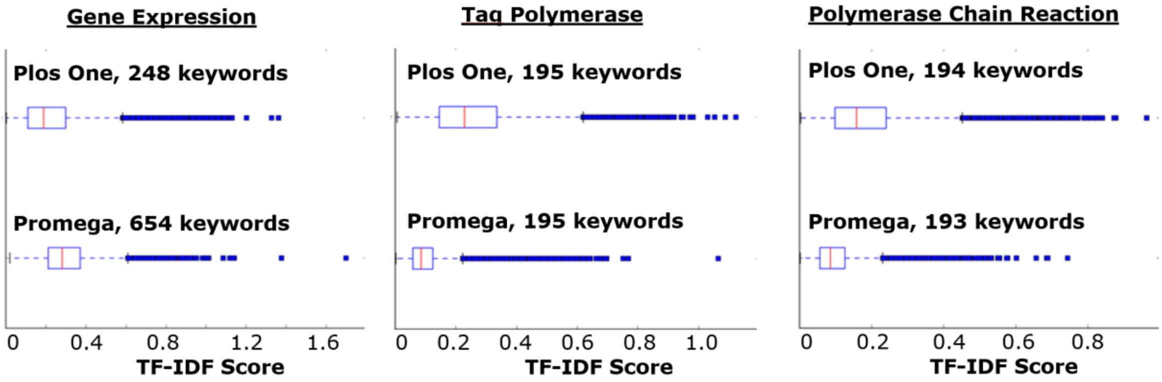


Figure 6 Comparison of scoring power between biotech company keywords and paper keywords.

There are a couple of observations from the plots above:

- Keywords harvested from PlosOne publication are comparable for Gene Expression products but they are much better for Tag Polymerase and Polymerase Chain Reaction products than those defined by the company's SMEs
- In the graph for Gene Expression, using more keywords doesn't produce higher scores: 248 keywords harvested from Plos One publications produce similar scores as 654 company-defined keywords.

The confidence level constructed in the previous section using Cell Signaling product only has been re-confirmed by company's SMEs to evaluate another 100 grant abstracts using harvested keywords from papers for each of the three products. A uniform stratification is used for this evaluation.

In terms of matched project counts with confidence level comparison below, we can see that using harvested PlosOne keywords produces more matched projects across all products, and they are 3 orders of magnitude better for Tag Polymerase and Ploymerase Chain Reaction products.

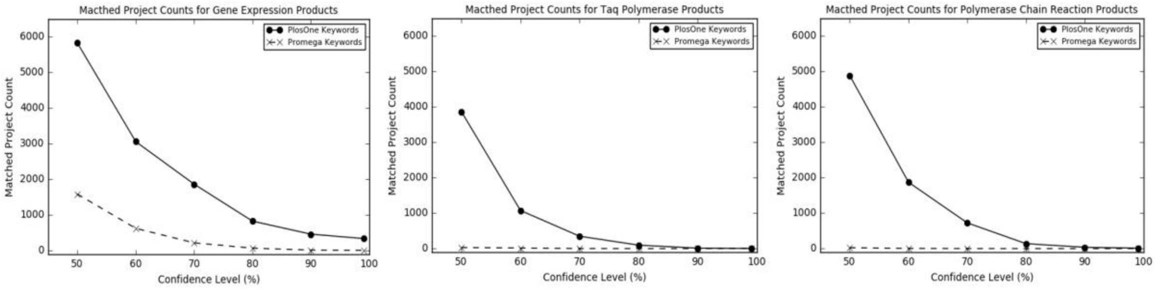


Figure 7 Comparison of matched projects between company keywords and paper keywords.

The scoring power in terms of matched projects using paper keywords are more comparable among products (see below). Obviously, the methodology with more uniform results is more desirable. Another key finding is that new applications and/or protocols which are unbeknownst to the company SMEs appears in the harvest keywords. In turn, the company can explore these newly discovered applications/protocols for marketing and sales campaigns.

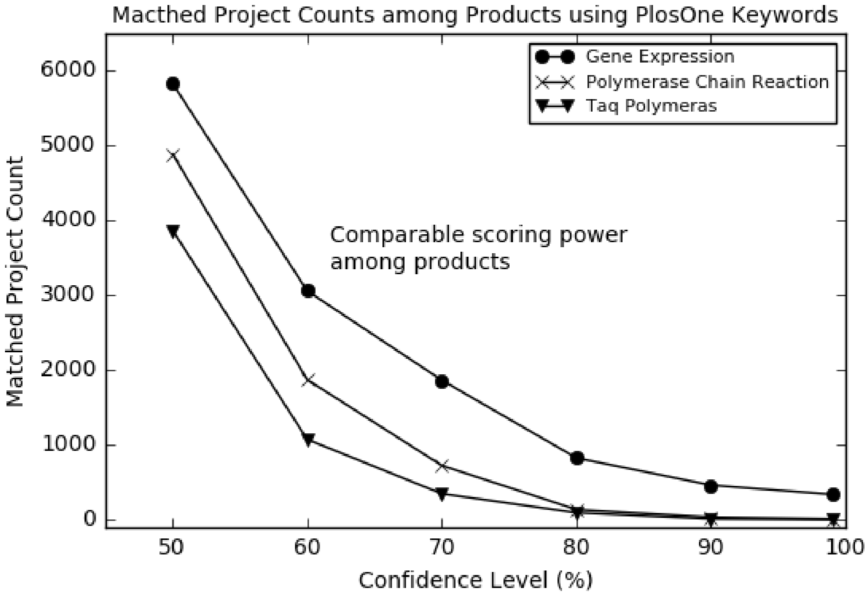


Figure 8 Comparison of matched projects among products using paper keywords.

6. Conclusions

We presented a methodology to support academic customer prospecting for biotech companies leveraging publicly available academic papers. This methodology uses text and graph analytics to give the biotech company visibility into who are their current academic customers, as well as provide leads for potential new customers and brand ambassadors within the academic community.

This methodology can be refined and expanded in many ways:

- Identification of biotech products used in a paper can be improved by use of supervised learning on a training set of papers with confirmed product association (through the use of SMEs). Such a set is not currently available, but it can be constructed as part of the biotech company's business processes. Each time a sales representative is presented with papers that the unsupervised algorithm from Section 3 associated with a product, they can provide feedback on whether that was a correct association or not. Over time, a training set of labeled papers would be developed, allowing for more powerful classification methods to be used.
- Identification of key influencers in product-specific research networks (Section 4) can be extended to monitor how authors move in the network through time, becoming more or less influential. Researchers in an ascending influence trend would be good potential prospects, even if they are not currently among the top influencers in the network.

- Establishing a method for identifying cross-selling opportunities based on biotech product sets that are associated to the same paper (i.e. are used in the same research project), using association pattern mining. Although this could already be done with current data, more reliable results would be obtained once the paper-product association is refined by supervised learning algorithms.

7. Bibliography

Aggarwal, C. C. (2015). *Data Mining: The Textbook.* Yorktown Heights, New York: Springer International Publishing Switzerland.

Apache. (n.d.). *Term frequency-inverse document frequency*. Retrieved from http://spark.apache.org/docs/latest/ml-features.html#tf-idf

Kovalchik, S. (2016, 11 2). *RISmed*. Retrieved 1 12, 2017, from https://cran.r-project.org/web/packages/RISmed/index.html

National Institutes of Health (NIH) . (n.d.). *ExPORTER download*. Retrieved 01 12, 2017, from https://exporter.nih.gov/ExPORTER_Catalog.aspx

National Institutes of Health (NIH) . (2016, April 4). Retrieved 01 12, 2017, from https://www.nih.gov/about-nih/what-we-do/budget

Scott Chamberlain, C. B. (2016, 11 24). *rplos*. Retrieved 1 12, 2017, from https://cran.r-project.org/web/packages/rplos/index.html

Wikipedia. (n.d.). *Term frequency-inverse document frequency*. Retrieved 01 12, 2017, from https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Dell EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL EMC MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.