Procedures for Electronic Analysis of Business Reports of 33 Afro-American Transnational Entrepreneurs

Decker, Arnim

Procedures for Electronic Analysis of Business Reports
of 33 Afro-American Transnational Entrepreneurs

Arnim Decker

Dept. International Business, Aalborg University, Denmark
decker@business.aau.dk

AALBORG UNIVERSITY DENMARK

Introduction

This is a feasibility study to test electronic text mining in the field of entrepreneurship research. Text Mining is useful for extracting high-quality information from text by devising patterns and trends. We created 3 types of representations: A) Dendrogram, B) Word cloud of most used words (here one document only), and C) bar chart of words with highest semantic significance (here one document only)

Shortended list of 33 reports

- Actuel Urban Livign PLC Business Plan.pdf
- African Business Expert Solutions - Rwanda Business Plan.pdf
- Associated Wind Developers, LLC Business Plan & Financial Plan.pdf
- BARRI-Industries Limited Business Plan & Financial Plan.pdf
- Biogen Kenya Business Plan & Financial Plan.pdf
- BisaDoc Business Plan & Financial Plan.pdf
- Brundo International PLC Business Plan.pdf
- etc....

Words cloud from one doc

Fig. 1: A word cloud extracted from one document in the sample (Actuel Urban Living PLC Business Plan)

Analysis

The sample was derived from from the ADM 2015 event (American Diaspora Marketplace). The US Federal Government undertakes a yearly business competition with the aim to enhance entrepreneurship in Africa. Entrepreneurs or entrepreneurial firms are invited to participate if they drive (or plan to drive) a business in Africa. To participate, participants must either hold an American passport, or be in possession of a Green Card. The motivation to participate is boosted by substatial price money, the winner of the competition wins 50.000 US$, while the 2nd and third can still earn 30.000 US$. In 2015, 400 entrepreneurial firms participated in the competition. Candidates delivered a business and a financial report and went through 3 elimination rounds in which intensive interviews are conducted, as well as an evaluation of the supplied reports. After 2 rounds, 33 entrepreneurs and entrepreneurial firms where still left in the competition, these participated in the last round which took place in Silversprings (Maryland) in Sept. 2015. As participants, we received and then analyzed the supplied 33 business reports electronically. The finalists are mainly active in agriculture and farming, other firms are dedicated to trading activities. There are also a few firms which were active in manufacturing, high tech, or engineering.

The electronic analysis: the aim was to electronically identify logical clusters to identify topic differences between the documents. For this purpose, we used the statistical programming language R. Our program proceeded in various steps. First the 33 pdf files where read into computer memory and some computational procedures where performed (cleaning for irrelevant words, tokenizing, etc). The individual words where then loaded into a computer memory representing a large matrix, where all words that appeared in the sample for 33 pdf files appear in rows (vertical axis). The names of the participating firms figure appear as columns (horizontal axis). To reduce the size of the matrix, *sparsity* was reduced meaning that cells which contain no information where eliminated. The program then applied the Ward.2 algorithm on the matrix. From the computed result, a dendrogramm was created (Fig.2). Apart from this we also created two graphs for each of the individual business reports. The first graphs represents a *word cloud* (Fig.1), with the most used words in the center of the cloud. The lesser the words in the report are used, the more they move to the outer limits of the cloud with correspondingly smaller fonts. This method gives a quick overview of the topic of a text. The other graph (Fig. 3) shows a bar chart, again one for each business report. This time the programm identified the semantically most important words and assigned a corresponding bar chart by semantic significance of the word.

Use of the Ward.2 algorithm for text analysis which is basically a way of calculating logical distance between objects for hierarchical cluster analysis,. This is the logical distance between the words in the above mentioned matrix. The result of the calculation is visualised in a dendrogram. It depicts a multilevel hierarchy in which higher positioned clusters are joined together and form smaller clusters at lower levels. This should allow for choices to identify logical clusters between the sample of documents.

Results: Fig. 1 and Fig. 3 demonstrate methods how to conveniently identify relevant terms for deeper analysis. Employing Ward.2 algorith for creating dendrogram proves to be problematic since the logic of the clustering is difficult to identify. As a next step in the research, we will **integrate a method for supervised analysis to gain more control over the logic for establishment of logical clusters**.

Dendrogram of the entire set of 33 documents

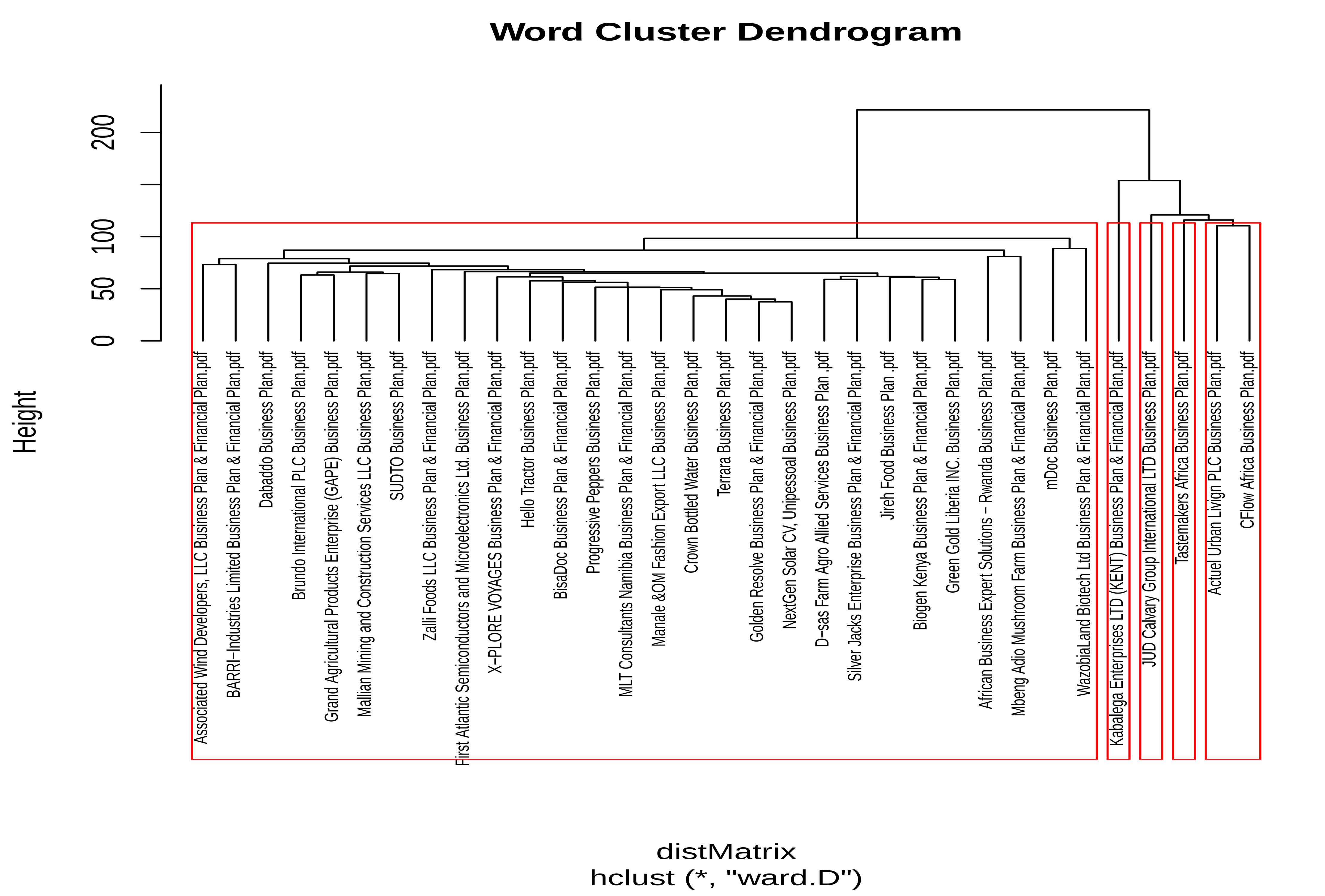

Fig. 2: A dendrogram drawn from the entire sample of 33 documents

Later todo list

- Improve and refine understanding of results of employed algorithm
- Use alternative algorithms
- Change parameter of dendrogram creation
- Use larger set of documents (test scalability)
- Move from untrained to trained methods of text analysis
- Test different preconditions for text clustering

Words from one doc

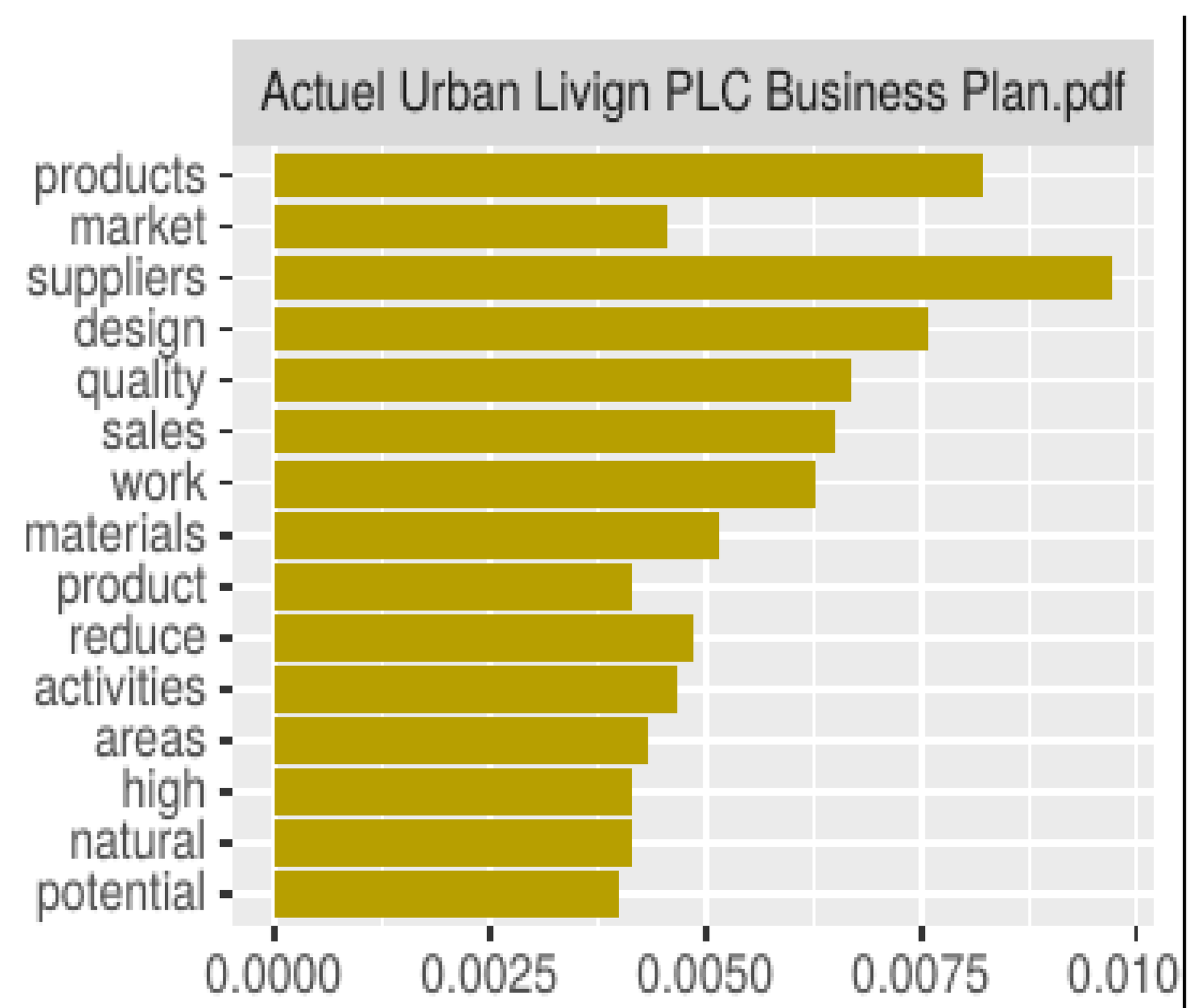

Fig. 3: A bar chart with semantically most significant words extracted from one document in the sample