Patent Classification and Prediction System

Data Science Institute
COLUMBIA UNIVERSITY

Raphael Kintzer, Jack Gilligan, Nae Eoun Lee, Alex Kong, Yufei Wu
Industry Mentor: Vikash Khanna
Faculty Mentor: Kriste Krstovski

Data Science Capstone Project with Unilever

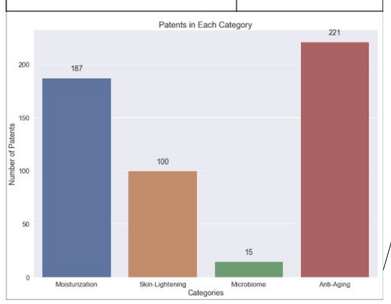
Project Goals:

To bolster competitor surveillance capabilities, Unilever asked us to develop the following tools to analyze skincare-related patents recently filed by their biggest competitor (L'Oreal):

- **A classifier to label patents as relating to any of four skincare categories: Anti-aging, Microbiome, Moisturization, and Skin Lightening**
- **Time series predictions for the distribution of future patents over the four categories**

Patents Data Set: Metadata

Date Range: Jan 2010 - Dec 2017	Total Patents: 997
Total Unlabeled Patents: 543	Mean Word Count: 827.2
Median word count: 771.0	Std. Word Count: 421.4



Final Patent Analysis System:

Patents

Key Features:
EPR: Patent filing date
Title: ~10 words, 1 language
Abstract: ~100 words; multiple languages
Claims: >500 words; multiple languages
IC Codes: Standard codes indicating to which technological area(s) patent pertains

Pre-processing

Classifier
- **Model:** Decision Tree classifier
- **Inputs:** Patent IC codes and all body text fields (vectorized using TF-IDF).
- **Outputs:** Binary labels for each of four skincare patent categories

Time Series Analysis
- **Model:** VAR (Vector AutoRegression)
- **Inputs:** Patent data with skincare category labels
- **Outputs:** Predicted patent category distributions for next *n* quarters.

Web App
- **Design:** Flask app to automate pre-processing, classification, and time series prediction (see demo).
- **Inputs:** Patent data, hyperparameters
- **Outputs:** Labeled patent data; plotted time series predictions

5 Most Common Words by Skincare Category

Most common words / # of occurrences	1st	2nd	3rd	4th	5th
All Patents	composition / 1816	comprising / 1307	cosmetic / 1120	relates / 940	skin / 765
Moisturization Patents	linking / 9	reaction / 8	resistant / 7	bonds / 6	layered / 6
Microbiome Patents	container / 6	longum / 6	bifidobacterium / 4	pro / 4	alterations / 4
Skin-lightening Patents	hyperpigmentation / 4	activator / 4	eo / 3	cb / 3	receptor / 3
Anti-Aging Patents	bundle / 9	trans / 8	filters / 8	threads / 8	photoprotective / 6

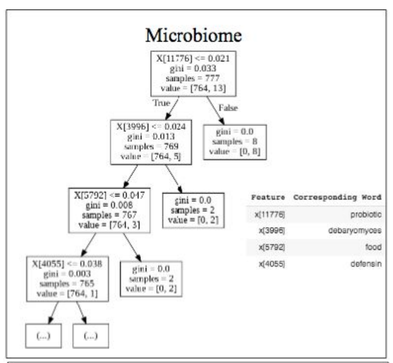
Patent Classifier:

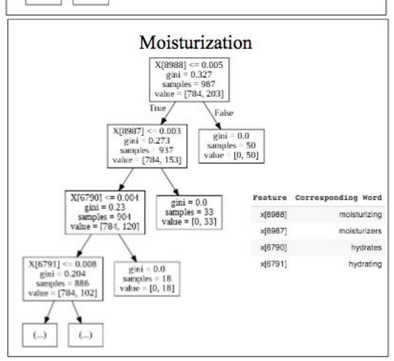
- **A pre-processing step cleans patent text, extracts its English translation only, and converts it to a TF-IDF vector.**
- **A patent's membership in the 'Anti-aging' and 'Skin Lightening' categories can be determined entirely from patent IC Codes, so no ML classifier is necessary.**
- **For 'Microbiome' and 'Moisturization' categories, we trained several models and compared their 5-fold cross-validated accuracies:**

Training validation results

Model	Cross-Validated Accuracy ('Moisturization')	Cross-Validated Accuracy ('Microbiome')
Naïve Bayes (Count Vectorizer)	74.52%	96.63%
Naïve Bayes (TF-IDF Vectorizer)	79.44%	98.50%
Decision Tree	**90.27%**	**98.51%**
Random Forest	81.14%	--
SVM	79.44%	--

Figure 1. Decision tree structure visualization

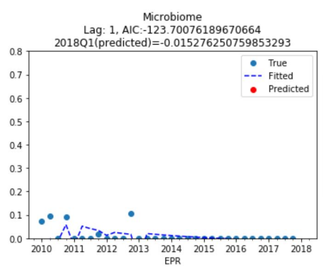




Patent Category Distribution - Time Series Predictions:

- **Labeled data is aggregated by quarter, creating 32 data points from 2010 - 2017.**
- **We fit several time series models to this data and compared AIC scores.**
- **A VAR multivariate model, which captures interactions between categories, achieved the best overall AIC score when forecasting all categories together.**
- **An ARIMA model, which captures non-seasonal patterns, was best for predicting individual category distributions.**

ARIMA(1,1,1) – Microbiome (lowest AIC among 5 categories)



AIC	-123.701
AR coeff. P>|z|	0.695
MA coeff. P>|z|	0.000

VAR – AIC and Correlation Matrix of Residuals

AIC: -346.818

	Moisturization	Microbiome	Skin Lightening	Anti-Aging	Others
Moisturization	1.000000	-0.337016	0.250984	-0.207905	-0.462162
Microbiome	-0.337016	1.000000	0.309548	-0.352126	-0.296629
Skin Lightening	0.250984	0.309548	1.000000	-0.358015	-0.721074
Anti-Aging	-0.207905	-0.352126	-0.358015	1.000000	0.625206
Others	-0.462162	-0.296629	-0.721074	0.625206	1.000000

Figure 2. (Left) ARIMA fit, AIC score, and significance of AR, MA terms. (Right) VAR model AIC score and correlation matrix between category distributions.

Web Application:

- **The user uploads a csv containing conforming patent data, as specified on the app's home page.**
- **The app classifies the data and returns a new csv with binary category indicator fields appended.**
- **The user can check this output, then upload it again to run time series analysis and obtain predictions as a plot and a table.**
- **Check out our demo application!**

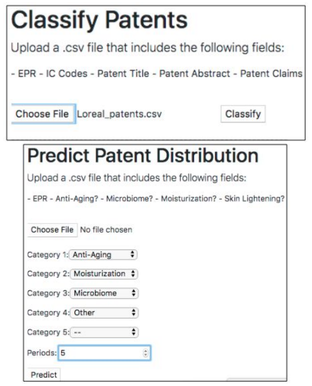


Figure 3. Web app homepage

Conclusion:

- **Our models, accessed via the web app, offer Unilever an intelligent way to track competitor patent activity.**
- **With more manually labeled training data examples, our classification and time series prediction models could potentially be improved.**
- **We will continue working with Unilever to develop additional visualization and analytical capabilities of the web app.**

Acknowledgments:

Special thanks to Kriste Krstovski and Vikash Khanna for their collaboration and support!

References:

[1] C. Bishop, *Pattern Recognition and Machine Learning*, Springer.

[2] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.