Module 8: NLP & Text Analysis

Data Science with Python – BSc Course

Why NLP & Text Analysis?

Over 300,000 financial news articles are published daily. Add earnings call transcripts, analyst reports, SEC filings, and social media — the volume of text is overwhelming. Bloomberg's NLP systems process millions of documents to extract trading signals. The firms that can read fastest, win.

In 2019, Thomson Reuters launched News Analytics powered by NLP. Two Sigma and Renaissance Technologies process alternative text data at scale. The message is clear: financial text is unstructured alpha waiting to be extracted.

Text data is the largest untapped signal source in modern finance

Why This Matters

- **News sentiment:** Thomson Reuters and Bloomberg provide NLP-driven trading signals from real-time news
- **Earnings analysis:** Automated earnings call transcript analysis predicts stock price reactions
- **Regulatory compliance:** Scanning millions of documents for AML/KYC patterns saves months of manual review
- **ESG scoring:** Extract sustainability signals from unstructured corporate reports and disclosures

Text processing skills unlock insights hidden in unstructured data

What You'll Learn

By the end of this module, you will be able to:

- Preprocess text data: tokenization, stemming, stopword removal
- Represent documents as numerical vectors using Bag-of-Words and TF-IDF
- Understand word embeddings and measure semantic similarity between words
- Build sentiment analysis models for financial text
- Apply NLP techniques to real financial documents

From raw text to actionable trading signals and insights

Lesson Roadmap

Lesson	Topic	Focus
L37	Text Preprocessing	Tokenization, stemming, stopwords
L38	Bag-of-Words and TF-IDF	Document vectors, term weighting
L39	Word Embeddings	Word2Vec, GloVe, semantic similarity
L40	Sentiment Analysis	Opinion mining, financial sentiment

Each lesson builds toward a complete NLP pipeline for finance

Key Concepts

- **Text Preprocessing Pipeline** – Clean and standardize raw text for analysis
- **Bag-of-Words & TF-IDF** – Convert documents into numerical feature vectors
- **Word Embeddings (Word2Vec, GloVe)** – Capture semantic meaning in vector space
- **Sentiment Analysis** – Extract positive/negative/neutral signals from financial text

Master text representation and you can mine insights from any document corpus

Finance Application

Scenario: Earnings Call Sentiment Analyzer
Using the skills from this module, you will build a system that:

- Preprocesses earnings call transcript text (tokenize, remove stopwords)
- Extracts TF-IDF features to identify key topics
- Uses word embeddings to capture management tone and semantic nuances
- Predicts stock price reaction based on earnings call sentiment scores

This is exactly what quantitative hedge funds do to generate alpha from alternative text data.

Turn unstructured earnings transcripts into structured trading signals

Who Uses This?

- **Data Providers:** Bloomberg and Refinitiv offer NLP-powered news analytics as core products
- **Hedge Funds:** Two Sigma processes news and social media for alpha; Renaissance uses text signals
- **Compliance:** HSBC and Deutsche Bank use automated document screening for AML/KYC compliance
- **ESG Analytics:** MSCI and Sustainalytics extract ESG signals from unstructured corporate reports

NLP is a competitive advantage across the entire financial ecosystem

What's Next: Module 9 – Deployment

You can build models. But a model in a notebook creates zero business value.
Module 9 teaches deployment: serializing models, exposing them via REST APIs, building interactive dashboards, and deploying to the cloud where they serve real users.
Prerequisite check: Can you preprocess text, build a TF-IDF vector, and train a sentiment classifier? If yes, you are ready to put it into production.

Models in production create value – Module 9 takes you from prototype to deployment

Let's Begin!

First up: L37 – Text Preprocessing

Open your laptop and follow along.