SPEECHMATICS

Metadata Generation PoC

Keyword and Topic Extraction - Solutions Brief

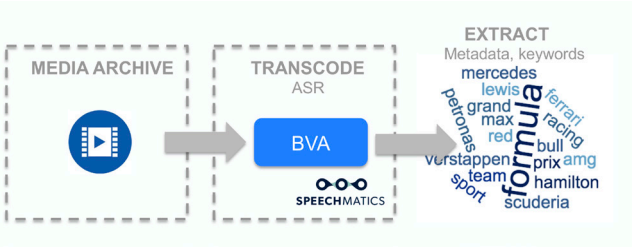
This Metadata Generation Proof of Concept demonstrates extraction of keywords and topics from text on a web page.

Metadata Extraction:

Massive amounts of digital content are driving the search for more efficient categorisation and metadata generation.

Using advanced machine learning techniques, Speechmatics is able to extract keywords and topics from an audio stream and integrate this, via a simple API, to a Media Asset Management (MAM) workflow.

This allows increasing volumes of content, whether VOD, adverts or audio clips, to be categorised. The monetisation potential of this content then increases, along with the operational efficiency of the content workflow.



Keywords and Topics:

Keywords are mentions of important terms and names in the audio stream - they allow you to quickly see the wood for the trees. Topics are terms that are not necessarily present in the transcript, but which are highly relevant for an understanding of the content.

For example, in a sports production workflow you might extract a set of topics from the speech stream in the digital audio and feed it into a recommendation engine.

Use Cases:

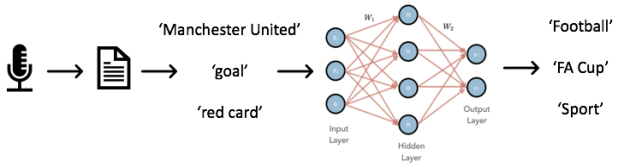
Newsrooms, marketing agencies and post-production teams will be interested in being able to extract relevant text from their stored media, either at ingest time or from an existing archive.

For example, if your interest is in creating topic digests for digital media, you can run Speechmatics' state-of-the-art ASR on the content to get an automated transcription, before generating topics from a set of keywords.

As part of your asset workflow the keywords and topics that exist in your digital content can be extracted, filtered, and made available to facilitate better search and recommendation of content.

Extraction using Machine Learning:

In order for the system to associate keywords with topics of interest, a hierarchy of topics is created from training data, which correlates topics to keywords.



Topic extraction implies that the machine learning algorithm has some knowledge of the domains that are of interest; for instance, it should know that Lewis Hamilton is a Formula 1 driver, and that he currently drives for the Mercedes AMG Petronas team.

The large amounts of speech training data that are available for Speechmatics to train with means that this knowledge is already captured for many of the most popular domains. This is an active area of research for Speechmatics as we look to improve and refine the ability of our topic extraction systems.

If you're interested in finding out more about how the media broadcast market is adopting Speechmatics' speech-to-text technology, and how we are working with some of our partners in this space, please visit our website:

https://www.speechmatics.com/sectors-and-solutions/media-broadcast/

Contact us for more information:

jamesp@speechmatics.com | +44 (0)1223 794 497 | **www.speechmatics.com**