Transformer-based embeddings vs Doc2Vec for skill–occupation analysis

1.1 Doc2Vec strengths:

- **Simple and lightweight.** Doc2Vec is computationally cheap and easy to train on large corpora of job postings.
- **No external resources required.** It learns directly from co-occurrence patterns in the data, without needing pretrained models.
- **Works reasonably well for broad patterns.** It can recover high-level occupational groupings when skill overlap is strong.

1.2 Doc2Vec limitations:

- **No semantic understanding.** Skills are represented the same way across all contexts, even when their meaning varies by occupation. (Example: if we embed two documents "Data Scientist - Python, SQL" and "Data Engineer - SQL, Python" using Doc2Vec, the two occupation vectors will be identical. However, if we first embed full definitions of the skills "Python" and "SQL" using a transformer model, "Data Scientist" will land closer to "Python" and "Data Engineer" to "SQL". This will help in the downstream task of ranking the skills by their importance for a particular occupation.)
- **Sensitive to frequency artifacts.** Common skills ("communication", "teamwork") dominate representations unless aggressively filtered.
- **Weak generalization to new or rare skills.** New skills must appear frequently in the training data to be meaningfully represented.

2.1 Transformer strengths:

- **Easier processing of a multilingual corpus.** Multilingual embedding models are trained so that sentences with the same meaning in different languages land in similar positions in vector space. That means we can embed text in English (from LightCast) and Italian (from the Regional Catalogue) with the same model and compare them directly with cosine similarity. No need to translate ex ante.
- **Contextual understanding of skills.** Transformers learn meaning from context, allowing the same skill to take different meanings depending on how relevant they are for different occupations. Embedding skill descriptions (either from LC taxonomy or the RegCal) will produce a vector that reflects the full meaning of the skill, not just the label token. Thus, skills with similar functions (like "neural networks" and "deep learning") will end up close in vector space even if they rarely co-occur. Additionally, skills can be embedded using: 1) skill descriptions, and 2) contexts extracted from job ads (if needed).
- **Zero-shot generalization to emerging skills.** New or emerging skills can be embedded immediately without retraining the entire model.

- **Explicit and interpretable aggregation.** Occupations can be constructed as weighted combinations of skill vectors, making skill importance more transparent.
- **Supports learning relevance, not just similarity.** With contrastive learning and attribution methods, transformers can help identify:
  - job-defining skills ("which skills distinguish an occupation", not just "which skills co-occur"),
  - redundant vs complementary skills,
  - realistic retraining requirements (estimating what skills a worker is actually missing, how important those skills are for the new profession, and how hard they are to acquire, rather than just saying that two jobs are "similar").

2.2 Transformer limitations:

- **Less "from-scratch"**
  Pretrained models embed general language knowledge, which may feel less transparent than training everything in-domain.
- **Overkill for very simple tasks**
  If the goal is only coarse clustering of occupations, the added complexity may not be necessary.

Doc2Vec provides a simple and computationally efficient baseline that can capture broad occupational patterns when skill overlap is strong. However, given the project's real context (multilingual data, short text inputs, a large and evolving skill taxonomy, and the need to identify not only similarity but also skill relevance and retraining requirements), its co-occurrence-based nature can be a limiting factor. **Transformer-based models** would be a better choice for embedding Italian and English text in a shared semantic space, capturing contextual skill meaning, and enabling transparent, skill-weighted occupation representations. Although they introduce additional complexity and rely on pre-trained language knowledge, this is compensated by their flexibility, potential for fine-tuning on job postings, and better alignment with the project's objectives.